3 PCA

3.1 Objective

The objective of the third chapter is to introduce principal component analysis, PCA. This is a multivariate projection method designed to extract and display the systematic variation in a data matrix X. A geometric perspective will be used to give an intuitive understanding of the method. With PCA a number of diagnostic and interpretational tools become available, which will be outlined.

3.2 Introduction to PCA

Principal component analysis forms the basis for multivariate data analysis [Jackson, 1991; Wold, *et al.*, 1984; Wold, *et al.*, 1987]. As shown by Figure 3.1, the starting point for PCA is a matrix of data with *N* rows (*observations*) and *K* columns (*variables*), here denoted by X. The observations can be analytical samples, chemical compounds or reactions, process time points of a continuous process, batches from a batch process, biological individuals, trials of a DOE-protocol, and so on. In order to characterize the properties of the observations one measures variables. These variables may be of spectral origin (NIR, NMR, IR, UV, X-ray, …), chromatographic origin (HPLC, GC, TLC, …), or they may be measurements from sensors in a process (temperatures, flows, pressures, curves, etc.).

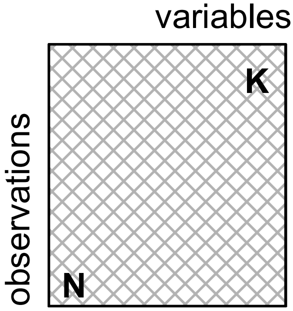


Figure 3.1: Notation used in PCA. The observations (rows) can be analytical samples, chemical compounds or reactions, process time points of a continuous process, batches from a batch process, biological individuals, trials of a DOE-protocol, and so on. The variables (columns) might be of spectral origin, of chromatographic origin, or be measurements from sensors and instruments in a process.

PCA goes back to Cauchy, but was first formulated in statistics by Pearson, who described the analysis as finding *lines and planes of closest fit to systems of points in space* [Jackson, 1991]. The most important use of PCA is indeed to represent a multivariate data table as a low-dimensional plane, usually consisting of 2 to 5 dimensions, such that an overview of the data is obtained. This overview may reveal groups of observations, trends, and outliers. This overview also uncovers the relationships between observations and variables, and among the variables themselves.

Statistically, PCA finds lines, planes and hyperplanes in the K-dimensional space that approximate the data as well as possible in the least squares sense. It is easy to see that a line or a plane that is the least squares approximation of a set of data points makes the variance of the co-ordinates on the line or plane as large as possible (Figure 3.2).

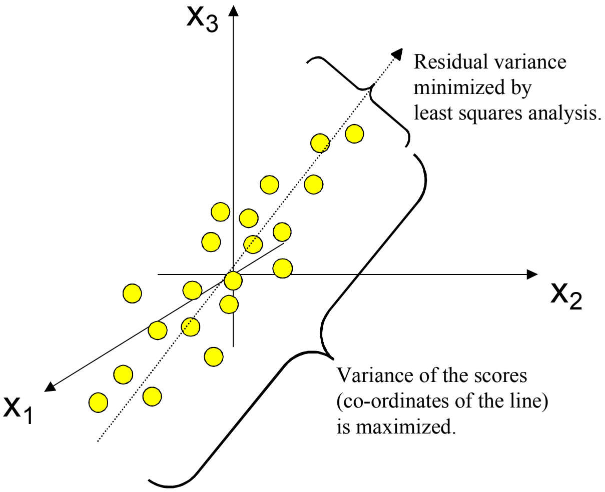


Figure 3.2: PCA derives a model that fits the data as well as possible in the least squares sense. Alternatively, PCA may be understood as maximizing the variance of the projection co-ordinates.

3.3 Pre-treatment of data

Prior to PCA, data are often pre-treated, in order to transform the data into a form suitable for analysis, i.e., to re-shape the data such that important assumptions are better fulfilled. In fact, pre-processing can make the difference between a useful model and no model at all. In this section *scaling of data* and *mean-centering* are described. Additional pre-processing tools like *transformations* (Chapter 9), *advanced scaling* (Chapter 10), and *data correction and compression* (Chapter 11) are addressed later.

3.3.1 Scaling

Variables often have substantially different numerical ranges. A variable with a large range has a large variance, whereas a variable with a small range has a small variance. Since PCA is a maximum variance projection method, it follows that a variable with a large variance is more likely to be expressed in the modelling than a low-variance variable. As an example, consider the LOWARP data set (Table 3.1), and particularly the wrp3 and st4 variables. Here, wrp3 varies between 0.2 and 1.0, whereas st4 ranges from roughly 17000 to 30000. As a consequence, st4 will dominate over wrp3, unless the data are normalized.

Table 3.1: The LOWARP data set.

		1	2	3	4	5	6	7	8	9	10	13	16	11	12	14	15	17	18
Num	Name	glas	crtp	mica	amtp	wrp1	wrp2	wrp3	wrp4	wrp5	wrp6	wrp7	wrp8	st1	st2	st3	st4	st5	st6
1	1	40	10	10	40	0.9	5	0.2	1	0.3	4.2	1.2	1.3	232	15120	2190	26390	2400	0.7
2	2	20	20	0	60	3.7	7.3	0.7	1.8	2.5	5.4	1.8	2.1	150	12230	905	20270	1020	0.6
3	3	40	20	0	40	3.6	6.9	0.9	2.1	4.8	9.4	1.2	1.4	243	15550	1740	21180	1640	
4	4	20	20	20	40	0.6	3.1	0.3	0.4	0.4	1.1	1	1	188	11080	1700	17630	1860	0.5
5	5	20	10	20	50	0.3	2.1	0.3	0.3	0.8	1.1	1.2	1.3	172	11960	1810	21070	1970	0.5
6	6	40	0	20	40	1.2	5					1.1	1.3	245	15600	2590	25310	2490	0.6
7	7	20	0	20	60	2.3	3.9	0.3	0.4	0.7	1.4	1.5	1.6	242	13900	1890	21370	1780	
8	8	40	0	10	50	2.6	5.9	0.4	0.2	0.7	1.2	1.6	1.6	243	17290	2130	30530	2320	0.7
9	9	30	20	10	40	2.2	5.3	0.2	0.7	0.6	2	1	1.1	204	11170	1670	19070	1890	0.6
10	10	40	0	0	60	5.8	7	0.9	1	5.6	11.8	1.6	1.8	262	20160	1930	29830	1890	
11	11	30	0	20	50	0.8	2.9	0.5	0.6	1.1	2	1.3	1.3	225	14140	2140	22850	2110	0.7
12	12	30	10	0	60	2.8	5.1	1	1.2	2.7	6.1	1.9	2.1	184	15170	1230	23400	1250	0.6
13	13	30	10	10	50	1.1	4.7	0.6	0.9	1.3	3.5	1.4	1.4	198	13420	1750	23790	1930	0.7
14	14	30	10	10	50	1.9	4.7	1	1	2.8	5.4	1.5	1.6	234	16970	1920	25010	1790	0.7
15	15	30	10	10	50	2.9	5.9	0.5	0.6	1	6.6	1.5	1.6	239	15480	1800	23140	1730	
16	16	40	10	0	50	5.5	7.9	0.8	2.4	5.5	9.3	1.5	1.8	256	18870	1880	28440	1790	
17	17	30	0	10	60	3.2	6	0.3	0.5	1.5	5.2	1.5	1.7	249	16310	1860	24710	1780	

Another simple example will further illustrate the concept of, and need for, scaling. In connection with a pre-season friendly game of football (soccer), the trainers of both teams decided to measure the body weight (in kg) of their players. The trainers also recorded the body height (in m) of each player. These data are listed in Table 3.2 and plotted in two ways in Figures 3.3 and 3.4.

Table 3.2: Measured body weights and body heights of 23 individuals.

Height (m)	1.8	1.61	1.68	1.75	1.74	1.67	1.72	1.98	1.92	1.7	1.77	1.92
Weight (kg)	86	74	73	84	79	78	80	96	90	80	86	93
Height (m)	1.6	1.85	1.87	1.94	1.89	1.89	1.86	1.78	1.75	1.8	1.68	
Weight (kg)	75	84	85	96	94	86	88	99	80	82	76	

When the two variables are plotted in a scatter plot where each axis has the same scale – the x and y axes both extend over 30 units – we can see that the data points *only* spread in the vertical direction (Figure 3.3). This is because body weight has a much larger numerical range than body height. Should we analyze these data with PCA, without any pre-processing, the results would only reflect the variation in body weight.

Actually, this data set contains an atypical observation (individual). This is much easier to see when the two variables are more appropriately scaled (Figure 3.4). Here, we have compressed the variation along the body weight axis and zoomed in on body height. There is a strong correlation between body height and body weight, except for one outlier in the data. This was impossible to see in the previous plot when body weight dominated over body height. We have therefore scaled the data such that both variables make the same contribution to the model.

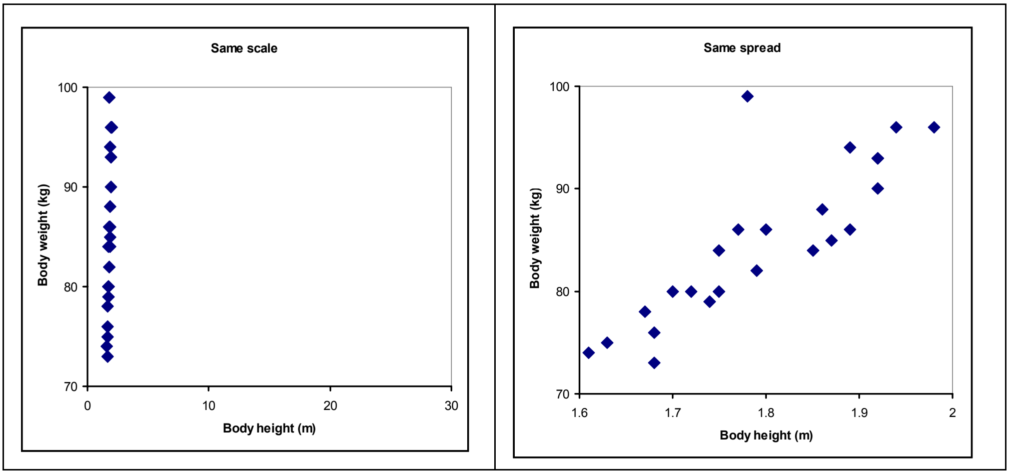


Figure 3.3: (left) Scatter plot of body weight versus body height of 23 individuals. The data pattern is dominated by the influence of body weight. The variables have the same scale. Figure 3.4: (right) Scatter plot of body weight against body height of 23 individuals. Now, the variables are given equal importance by displaying them according to the same spread. An outlier, a deviating individual, is now discernible. We have good reasons to believe (admittedly, based on some detailed insight of one of the Umetrics' staff members) that the outlying observation (the atypical person) happens to be the referee of the game.

In order to give both variables, body weight and body height, equal weight in the data analysis, we standardized them. Such a standardization is also known as "scaling" or "weighting", and means that the length of each co-ordinate axis in the variable space is regulated according to a pre-determined criterion (Figure 3.5). The first time a data set is analyzed it is recommended to set the length of each variable axis to equal length.

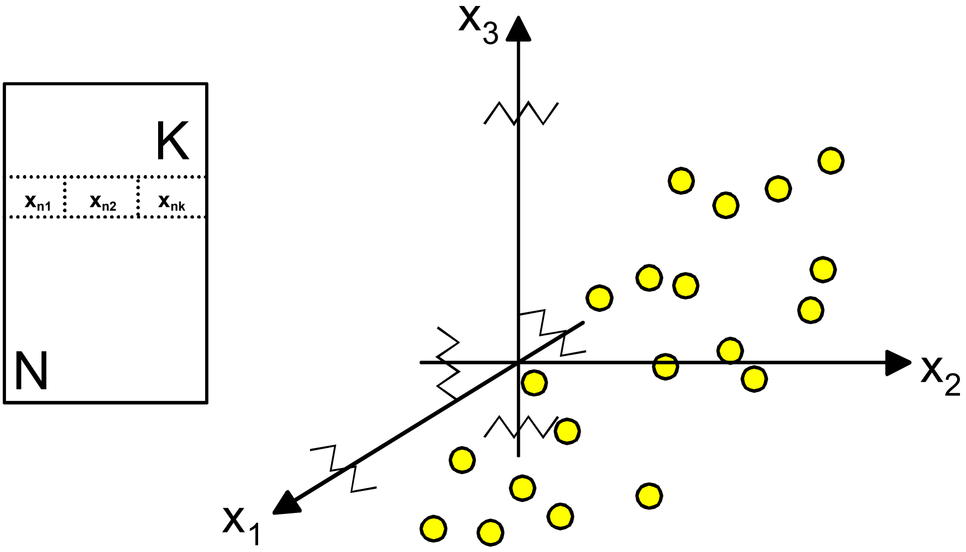


Figure 3.5: The scaling of variables means that the length of each co-ordinate axis in the variable space is regulated according to some selected criterion. The most common criterion is that the length of each variable axis be set to be the same variance.

There are many ways to scale the data, but the most common technique is the *unit variance (UV) scaling*. For each variable (column) one calculates the standard deviation (s_k) and obtains the scaling weight as the inverse standard deviation (*$1/s_k$*). Subsequently, each column of **X** is multiplied by *$1/s_k$*. Each scaled variable then has equal (unit) variance. Another name for this scaling method is “auto-scaling”.

A simple geometrical understanding of UV-scaling is based on the equivalence between the length of a vector and its standard deviation (square root of variance). Hence, the initial variance of a variable is interpretable as the squared “size” or “length” of that variable. This means that with UV-scaling we accomplish a shrinking of “long” variables and a stretching of “short” ones (Figure 3.6). By putting all variables on a comparable footing, no variable is allowed to dominate over another because of its length.

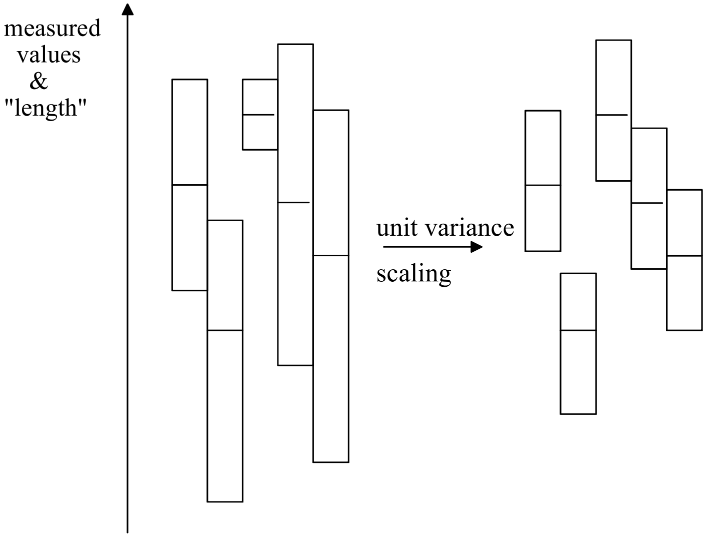


Figure 3.6: The effect of unit variance scaling. The vertical axis represents the “length” of the variables and their numerical values. Each bar corresponds to one variable and the short horizontal line inside each bar represents the mean value. Prior to any pre-processing the variables have different variances and mean values. After scaling to unit variance, the “length” of each variable is identical. The mean values still remain different, however.

Like any projection method PCA is sensitive to scaling. This means that by modifying the variance of the variables, it is possible to attribute different importance to them. This gives the possibility of down-weighting irrelevant or noisy variables. However, one must not overlook the risk of scaling subjectively to give you the model you want. Generally, UV-scaling is the most objective approach, and is recommended if there is no prior information about the data. Sometimes no scaling at all would be appropriate, especially with data where all the variables are expressed in the same unit, for instance, with spectroscopic data. Later on, when more experience has been gained, more elaborate scaling procedures may be used.

3.3.2 Mean-centering

Mean-centering is the second part of the standard procedure for pre-preprocessing. With mean-centering the average value of each variable is calculated and then subtracted from the data. This improves the interpretability of the model. A graphical interpretation of mean-centering is shown in Figure 3.7.

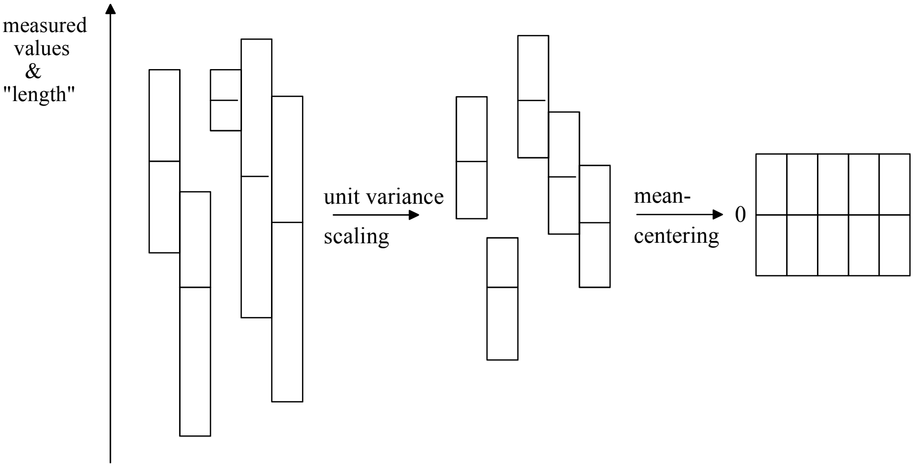


Figure 3.7: After mean-centering and unit variance scaling all variables will have equal "length" and mean value zero.

The mean-centering and UV-scaling procedures are applied by default in SIMCA-P. Note, however, that in some cases, such as multivariate calibration, it is not necessarily advantageous to use this combination of pre-processing tools, and some other choice might be more appropriate.

3.4 A geometric interpretation of PCA

We will now explain how PCA works: initially, using a geometrical approach, followed by a more formal algebraic account.

3.4.1 Setting up K-dimensional space

Consider a matrix **X** with *N* observations and *K* variables. For this matrix we construct a variable space with as many dimensions as there are variables (Figure 3.8). Each variable represents one co-ordinate axis. For each variable the length has been standardized according to a scaling criterion, normally by scaling to unit variance.

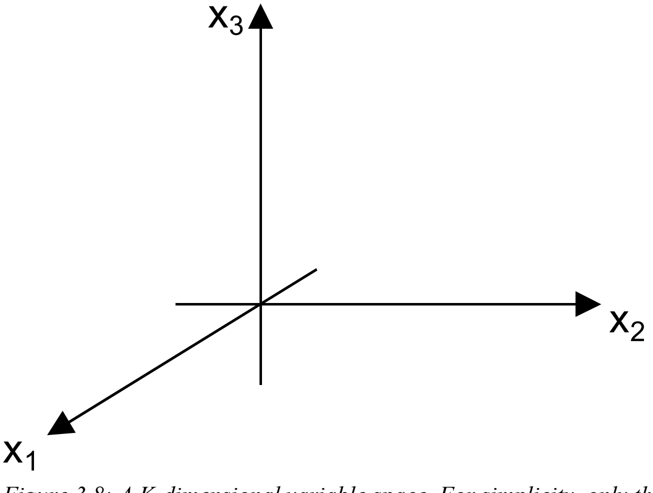


Figure 3.8: A K-dimensional variable space. For simplicity, only three variable axes are displayed. The "length" of each co-ordinate axis has been standardized according to a specific criterion, usually unit variance scaling.

3.4.2 Plotting the observations in K-dimensional space

In the next step, each observation (each row) of the X-matrix is placed in the K-dimensional variable space. Consequently, the rows in the data table form a swarm of points in this space (Figure 3.9).

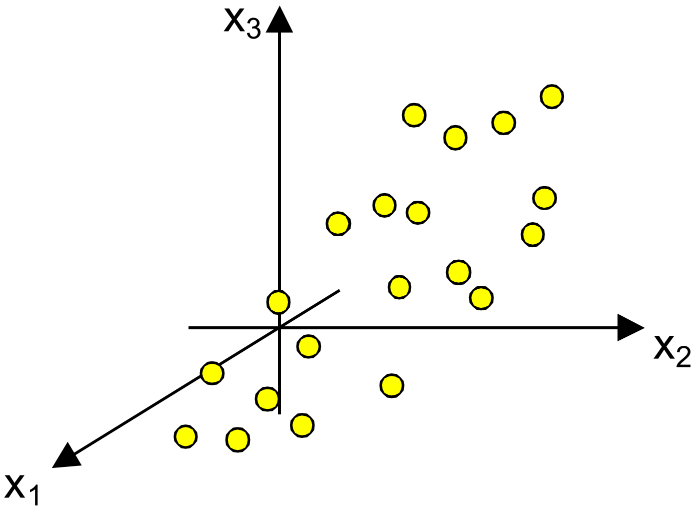


Figure 3.9: The observations (rows) in the data matrix X can be understood as a swarm of points in the variable space (K-space).

3.4.3 The effect of mean-centering

The mean-centering involves the subtraction of the variable averages from the data. This vector of averages corresponds to a point in the K-space (Figure 3.10).

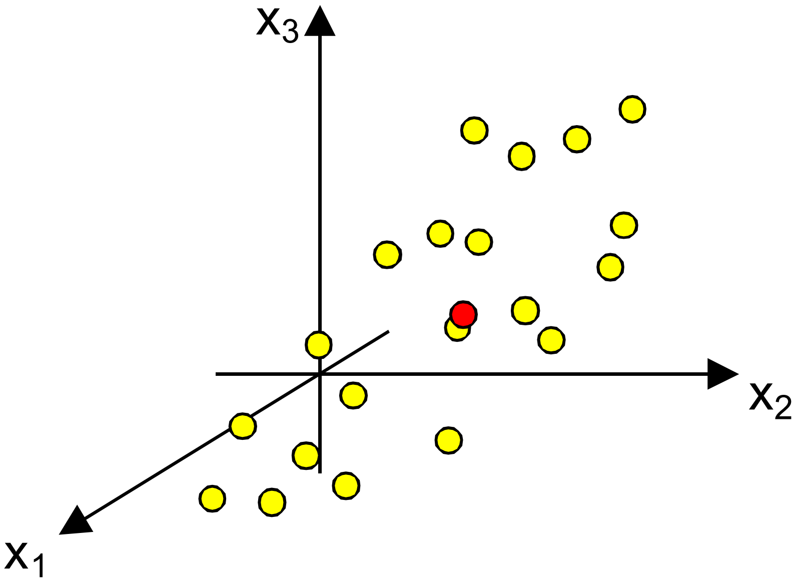


Figure 3.10: In the mean-centering procedure one first computes the variable averages. This vector of averages is interpretable as a point (here: in dark gray) in space. This point is situated in the middle of the point swarm (at the center of gravity).

The subtraction of the averages from the data corresponds to a re-positioning of the co-ordinate system, such that the average point now is the origin (Figure 3.11).

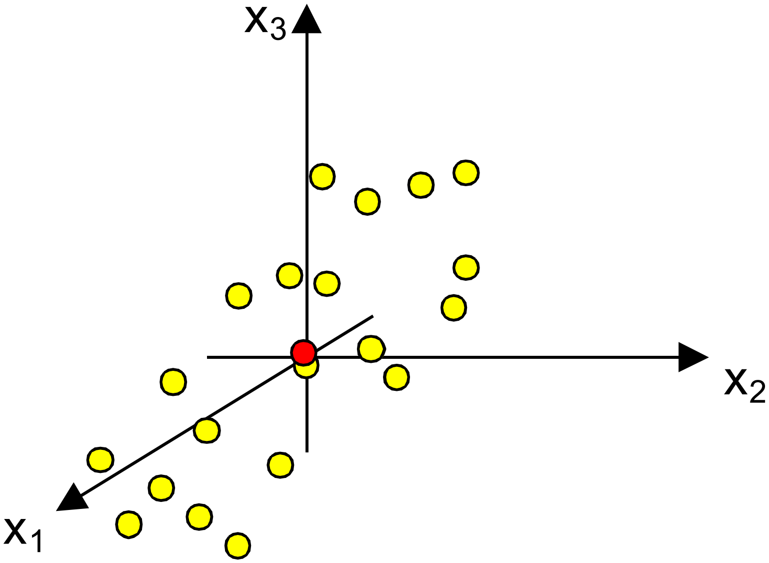


Figure 3.11: The mean-centering procedure corresponds to moving the origin of the co-ordinate system to coincide with the average point (here: in dark gray).

3.4.4 The first principal component

After mean-centering and scaling to unit variance the data set is ready for the computation of the first principal component (PC1). This component is the line in the K-dimensional space that best approximates the data in the least squares sense. This line goes through the average point (Figure 3.12). Each observation may now be projected onto this line in order to get a co-ordinate value along the PC-line. This new co-ordinate value is known as a *score*.

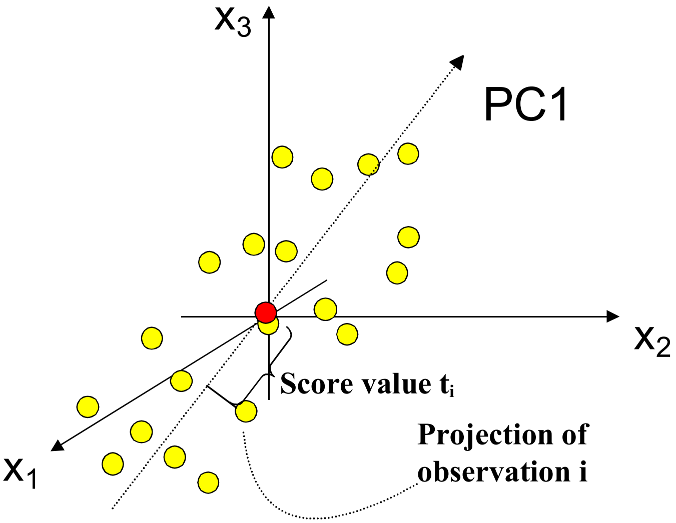


Figure 3.12: The first principal component, PC1, is the line which best accounts for the shape of the point swarm. It represents the maximum variance direction in the data. Each observation may be projected onto this line in order to get a co-ordinate value along the PC-line. This value is known as a score.

3.4.5 Extending the model with the second principal component

Usually, one principal component is insufficient to model the systematic variation of a data set. Thus, a second principal component, PC2, is calculated. The second PC is also represented by a line in the K-dimensional variable space, which is orthogonal to the first PC (Figure 3.13). This line also passes through the average point, and improves the approximation of the X-data as much as possible.

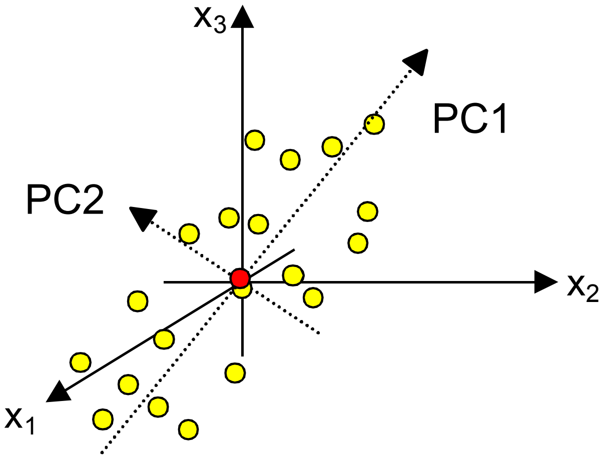


Figure 3.13: The second principal component, PC2, is oriented such that it reflects the second largest source of variation in the data, while being orthogonal to the first PC. PC2 also passes through the average point.

3.4.6 Two principal components define a model plane

When two principal components have been derived they together define a plane, a window into the K-dimensional variable space (Figure 3.14). By projecting all the observations onto this low-dimensional sub-space and plotting the results, it is possible to visualize the structure of the investigated data set. The co-ordinate values of the observations on this plane are called *scores*, and hence the plotting of such a projected configuration is known as a *score plot*.

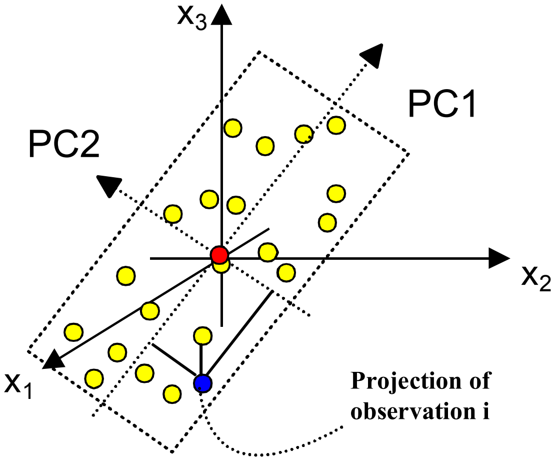


Figure 3.14: Two PCs form a plane. This plane is a window into the multidimensional space, which can be visualized graphically. Each observation may be projected onto this giving a score for each.

Let us now reconsider the FOODS data set. Figure 3.15 displays the plot obtained when plotting the scores of the two first principal components. These scores are called t_1 and t_2 (this notation is better explained in section 3.4.9).

As seen in Figure 3.15, each European country is characterized by two values, one along the first PC and another along the second PC. This score plot is a map of the 16 countries. Countries close to each other have similar properties, whereas those far from each other are dissimilar with respect to food consumption profiles. The Nordic countries (Finland, Norway, Denmark and Sweden) are located together in the upper right-hand corner, thus representing a group of nations with some similarity in food consumption. Belgium and Germany are close to the center (origin) of the plane, which indicates that they have average properties.

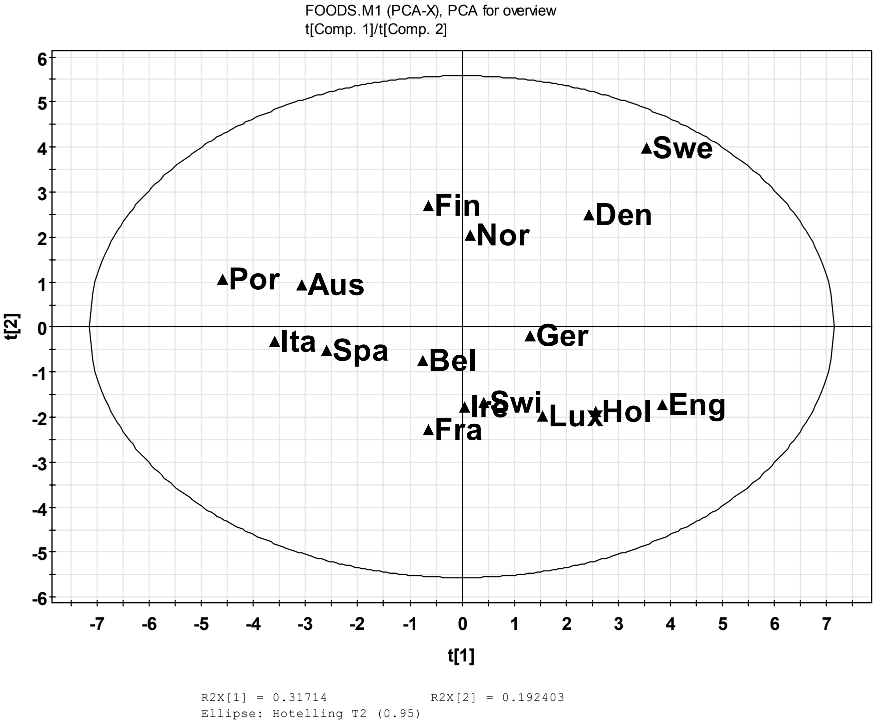


Figure 3.15: PCA score plot of the two first PCs of the FOODS data set. This provides a map of how the countries relate to each other. The first component explains 32% of the variation and the second component 19%.

3.4.7 How to interpret the score plot

In a PCA model with two components, that is, a plane in K-space, we wonder which variables are responsible for the patterns seen among the observations? We would like to know which variables are influential, and also how the variables are correlated. Such knowledge is given by the principal component *loadings* (Figure 3.16). These loading vectors are called p_1 and p_2 (see further discussion in Section 3.4.9).

Figure 3.16 displays the relationships between *all* 20 variables at the same time. Variables contributing similar information are grouped together, that is, they are correlated. Crisp bread (Crisp_Br) and frozen fish (Fro_Fish) are examples of two variables which are positively correlated. When the numerical value of one variable increases or decreases, the numerical value of the other variable has a tendency to change in the same way.

When variables are negatively ("inversely") correlated they are positioned on opposite sides of the plot origin, in diagonally opposed quadrants. For instance, the variables garlic and sweetener are inversely correlated, meaning that when garlic increases sweetener decreases, and vice versa.

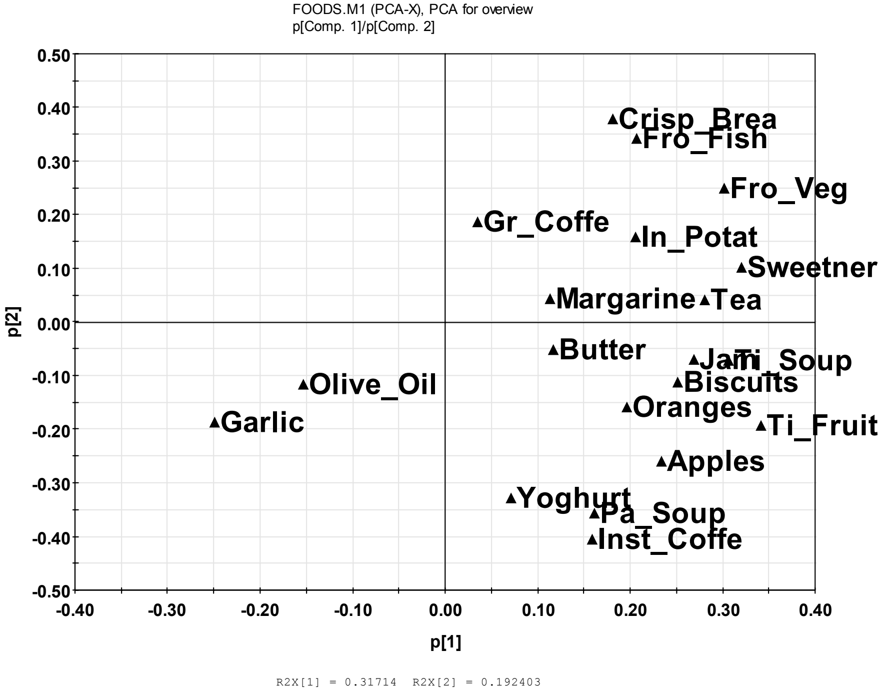


Figure 3.16: PCA loading plot of the first two principal components (p_2 vs. p_1) of the FOODS data set.

Furthermore, the *distance* to the origin also conveys information. The further away from the plot origin a variable lies, the stronger impact that variable has on the model. This means, for instance, that the variables crisp bread (Crisp_Br), frozen fish (Fro_Fish), frozen vegetables (Fro_Veg) and garlic (Garlic) separate the four Nordic countries from the others. The four Nordic countries are characterized by having high values (high consumption) of the former three provisions and low consumption of garlic. Moreover, the model interpretation suggests that countries like Italy, Portugal, Spain, and to some extent Austria, have high consumption of Garlic, and low consumption of sweetener (Sweetener), tinned soup (Ti_Soup) and tinned fruit (Ti_Fruit).

Geometrically, the principal component loadings express the orientation of the model plane in the K-dimensional variable space (Figure 3.17). The direction of PC1 in relation to the original variables is given by the cosine of the angles α_1, α_2, and α_3. These values indicate how the original variables x_1, x_2, and x_3 "load" into (= contribute to) PC1. Hence, they are called loadings. Of course, a second set of loading coefficients expresses the direction of PC2 in relation to the original variables. Hence, with two PCs and three original variables, six loading values (cosine of angles) are needed to specify how the model plane is positioned in the K-space.

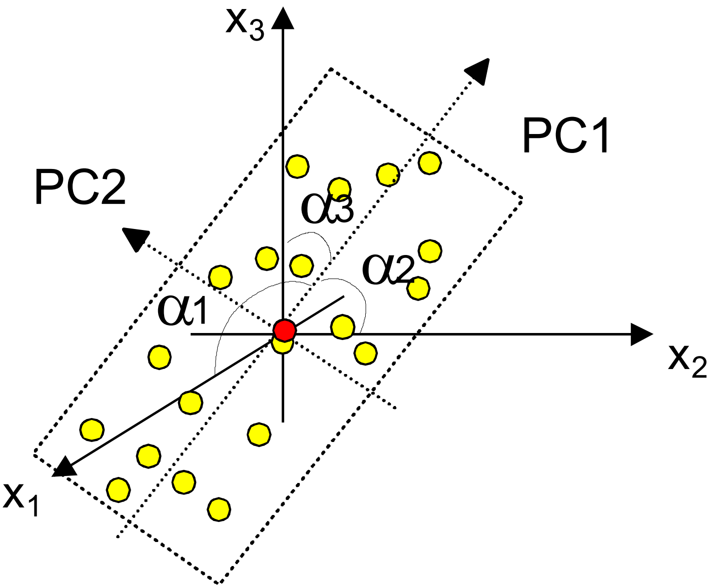


Figure 3.17: The principal component loadings uncover how the PCA model plane is inserted in the variable space. The loadings are used for interpreting the meaning of the scores.

3.4.8 Extensions to higher-order components

Frequently, one or two principal components are not enough to adequately summarize the information in a data set. In such cases, the descriptive ability of the PCA model improves by using more principal components. There are several approaches that can be used to evaluate how many principal components are appropriate [Jackson, 1991].

Consider the two-dimensional PCA model of the FOODS data set. We computed a third principal component. This third PC (i) is oriented in the direction of the third largest variation in the data, (ii) is orthogonal to the other two, and (iii) passes through the average point (the origin). The orthogonality constraint thus means that the third component becomes perpendicular to the already existing model plane.

The scores of the third and first PCs (t_3 vs. t_1) are plotted in Figure 3.18. Thus, in comparison with Figure 3.15, the vertical axis has been changed, while the horizontal axis is preserved. The most striking feature in this score plot (Figure 3.18) is the deviating behavior of England and Ireland.

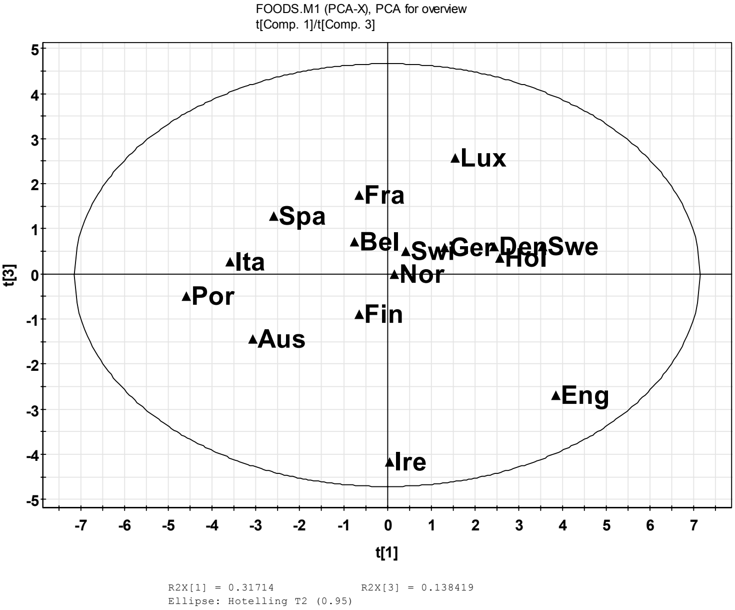


Figure 3.18: PCA score plot of the third PC (t_3) versus the first PC (t_1) of the FOODS data set. The first component explains 32% of the variation and the third component 14%.

The scores of the third PC are accompanied by the corresponding loadings. A scatter plot of the loadings of the third component versus the loadings of the first component (p_3 vs. p_1) is shown in Figure 3.19. This plot indicates that it is mainly the variables tea and jam which govern the positioning of England and Ireland. Especially England, but to some extent also Ireland, consumes larger amounts of Tea and Jam compared with the other countries included in this survey. Additionally, England and Ireland exhibit less than average consumption of garlic and olive oil.

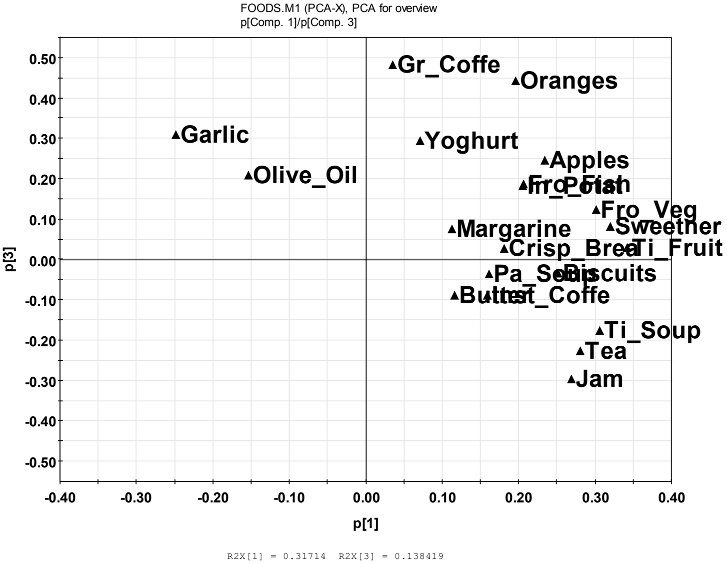


Figure 3.19: PCA loading plot of the loadings of the third and first principal components (p_3 vs. p_1) of the FOODS data set.

3.4.9 Summary of PCA

By using PCA a data table **X** is modelled as

$$\mathbf{X} = \mathbf{1} * \bar{\mathbf{x}}' + \mathbf{T} * \mathbf{P}' + \mathbf{E} \qquad \text{(eqn. 3.1)}$$

In the expression above, the first term, $\mathbf{1} * \bar{\mathbf{x}}'$, represents the variable averages and originates from the pre-processing step. The second term, the matrix product **T*P´**, models the structure, and the third term, the residual matrix **E**, contains the noise.

The principal component scores of the first, second, third, …, components ($\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$, …) are columns of the score matrix **T**. These scores are the co-ordinates of the observations in the model (hyper-)plane. Alternatively, these scores may be seen as new variables which summarize the old ones (Figure 3.20). In their derivation, the scores are sorted in descending importance (t_1 explains more variation than t_2, t_2 explains more variation than t_3, and so on). Typically, 2 to 5 principal components are sufficient to approximate a data table well.

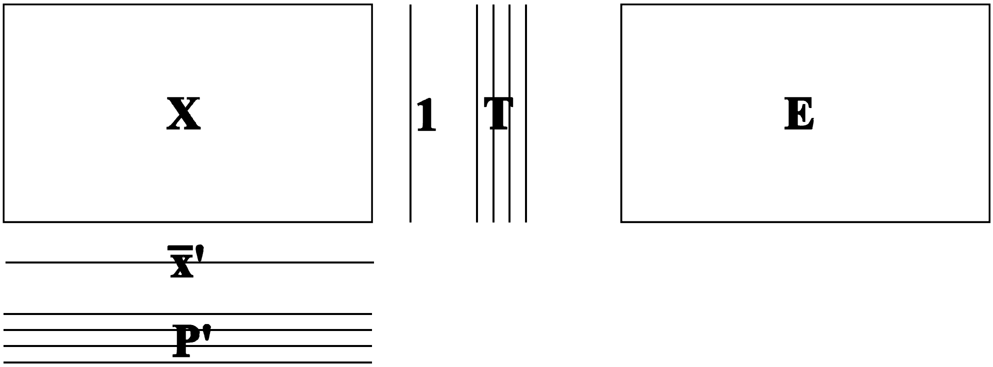


Figure 3.20: A matrix representation of how a data table X is modelled by PCA.

The meaning of the scores is given by the loadings. The loadings of the first, second, third, …, components ($\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{p}_3$,..) build up the loading matrix **P** (Figure 3.20). Note that in Figure 3.20 a prime has been used with **P** to denote its transpose.

The loadings define the orientation of the PC plane with respect to the original X-variables. Algebraically, the loadings inform how the variables are linearly combined to form the scores. The loadings unravel the *magnitude* (large or small correlation) and the *manner* (positive or negative correlation) in which the measured variables contribute to the scores.

To clarify what is plotted and when, Figure 3.21 provides a final overview of the various scores and loadings plots pertaining to the FOODS example.

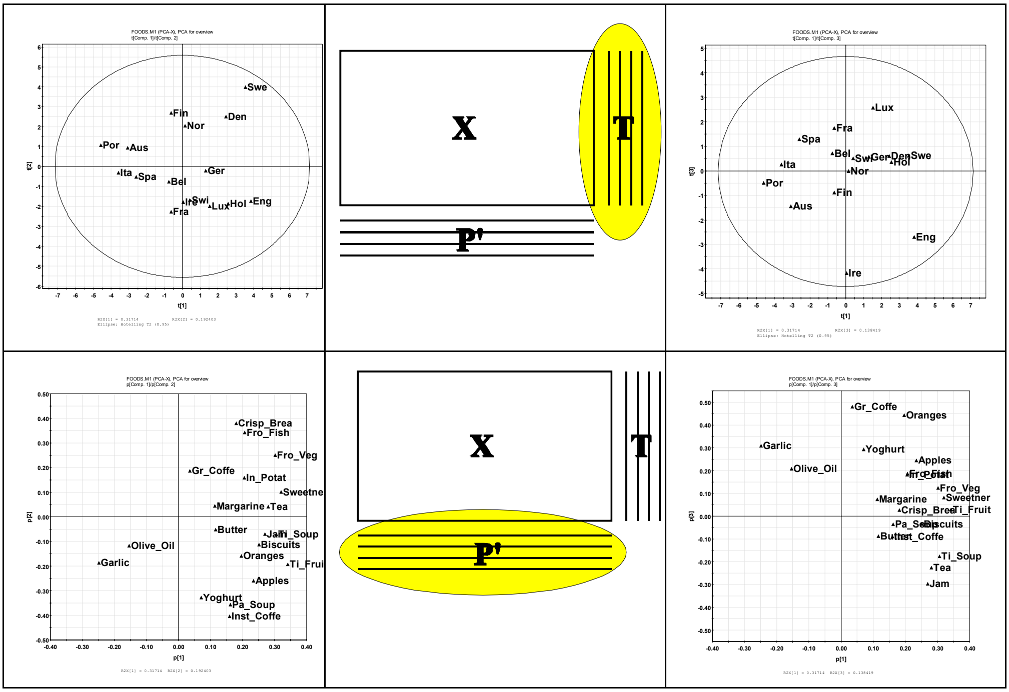


Figure 3.21: An overview of which PCA parameters are plotted in connection with the FOODS data set.

3.5 Additional PCA diagnostics

PCA offers a number of useful model parameters and other diagnostic tools, which can be displayed graphically or numerically [Wold, *et al.*, 1984; Wold, *et al.*, 1987]. In this section, we will continue the discussion of scores and loadings, but also discuss the *residuals*, the deviations of the real data from the model, and explore some diagnostics related to these. In addition, the technique of cross-validation will be outlined; this assesses the complexity and predictive power of the model.

3.5.1 Observation diagnostics – Are there outliers in the data?

PCA discovers *strong* outliers and *moderate* outliers. Conceptually, outliers are observations that are extreme or that do not fit the PCA model. Outliers are both serious and interesting, but easy to detect. Strong outliers are found in plots of PCA scores and moderate outliers are found by inspecting the model residuals [Wikström, *et al.*, 1998a]. By the term *residuals* we mean the X-variation that was not captured by the PCA model, the variation which constitutes the matrix **E** in equation 3.1.

Strong outliers are found in the score plots. They have high *leverage* on the model, i.e., strong "power" to pull the PCA model toward themselves, and may "consume" one PC just because of their existence (Figures 3.22 and 3.23). The term *leverage* derives from the Archimedean principle that anything can be lifted out of balance as long as the lifter has a long enough lever. Leverage is a measure of the influence of an observation and is proportional to the distance of the observation from the center of the data (see details in Appendix II).

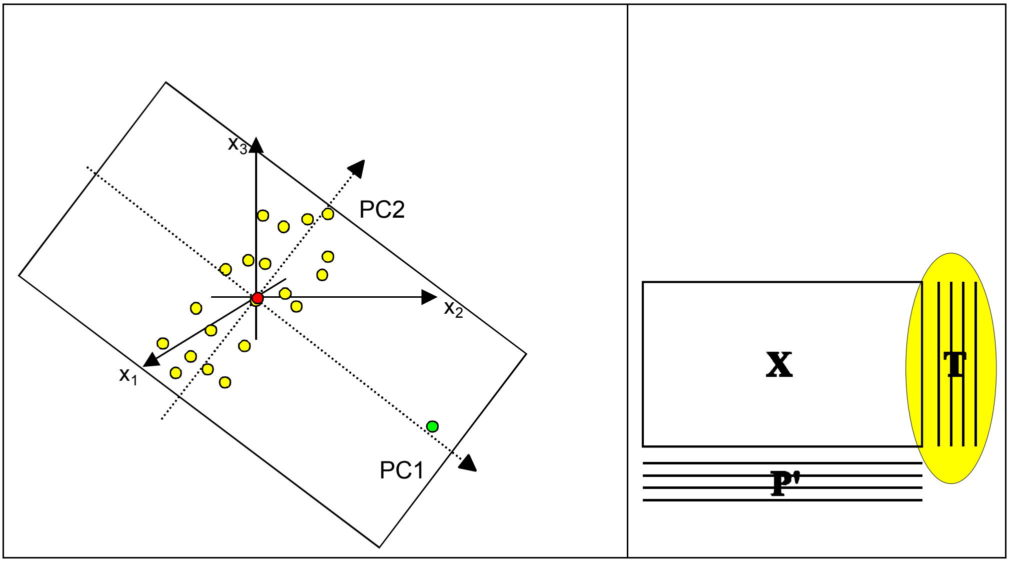


Figure 3.22: (left) Strong outliers are found in score plots. They have high leverage *on the model, i.e., strong "power" to rotate the PCA model towards themselves. Figure 3.23: (right) Plotting of PCA scores is useful for identifying strong outliers.*

A diagnostic showing strong outliers is given by *Hotelling's* T^2 [Jackson, 1991; Wikström, *et al.*, 1998a]. This statistic is a multivariate generalization of Student's t-test, and provides a check for observations adhering to multivariate normality. A definition of Hotelling's T^2 is given in Appendix II.

When used in conjunction with a score plot, Hotelling's T^2 defines the normal (operating) area corresponding to, for instance, 95% or 99% confidence. Figures 3.24 and 3.25 demonstrate cases where strong outliers are present and absent, respectively.

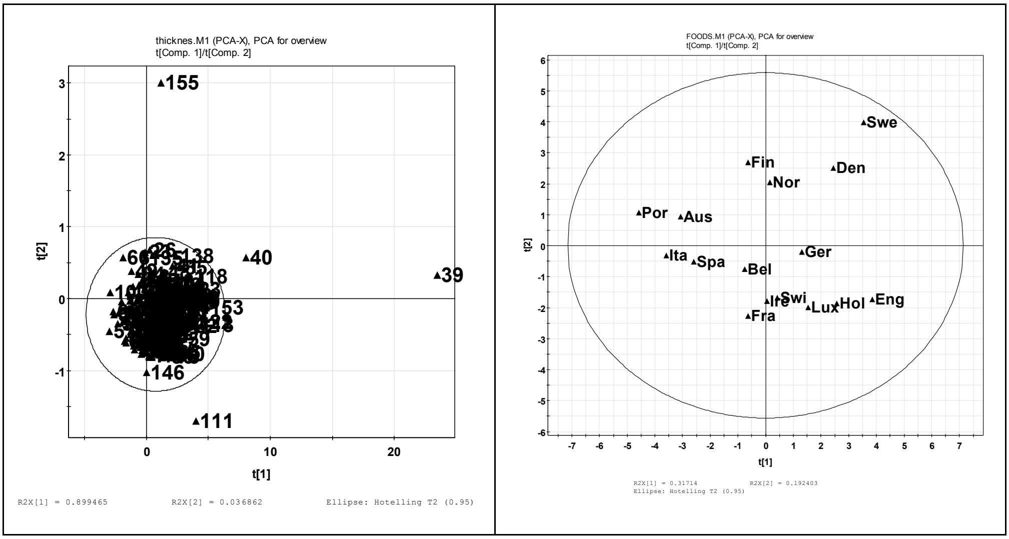


*Figure 3.24: (left) PCA score plot of a process. Observation 39 is a very strong outlier and has shifted PC1 with respect to itself. Also observations 40, 111, and 155 fall clearly outside the confidence ellipse. Figure 3.25: (right) PCA score plot of the FOODS data set. This data set contains no strong outlier. Recall that with 16 observations around 16*0.05 = 0.8 observation is expected to be outside the Hotelling's* T^2 *tolerance ellipse.*

In Figure 3.24, the extreme character of observation number 39 is beyond all doubt. This sample, and perhaps also samples 40, 111 and 155, ought to be more closely inspected. In order to better resolve and examine the main cluster, the four indicated samples must be temporarily omitted and a new PCA model fitted. Notice, however, that an outlier of this sort (cf. #39) may well be the interesting case, which should be looked at in detail in future investigations.

Figure 3.25 shows the first score plot of the FOODS data set. Here, no strong outliers are seen. The data points are fairly evenly scattered and it is easy to overview the relationships among them. In addition, it must be stressed that with *N* observations, it is to be expected that around N*0.05 observations will be found outside the 95% confidence region. Only a handful of these “potential” outliers are likely to be “real” outliers. Hence, in any modelling, insight into the process is required to sort out the “real” ones.

A data set may also contain moderate outliers, which are not powerful enough to shift the model plane and hence show up as outliers in score plots. Moderate outliers are identified by the residuals of each observation. In SIMCA-P, the detection tool for moderate outliers is called DModX, a short-hand notation for *distance to the model in X-space* (Figure 3.26). DModX is based on considering the elements of the residual matrix **E** and summarizing these row-by-row (Figure 3.27).

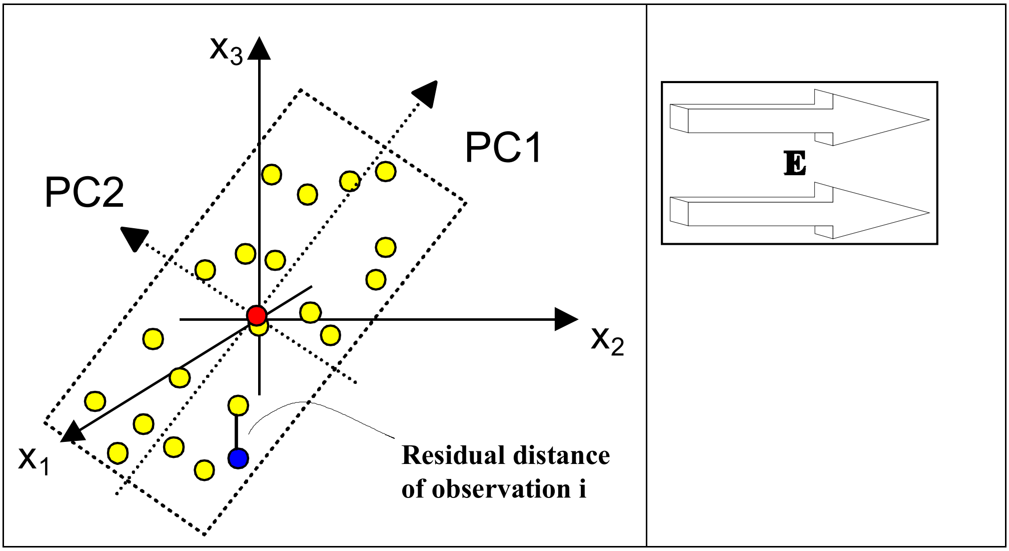


Figure 3.26: (left) A geometrical interpretation of an observation’s distance to the model (DModX). A value for DModX can be calculated for each observation and these values may be plotted against the time order or observation number together with a typical deviation distance (Dcrit) in order to reveal moderate outliers. Figure 3.27: (right) In the computation of DModX, the residuals of the matrix E are summarized row-by-row.

A value for DModX can be calculated for each observation. These values can be plotted in a control chart where the maximum tolerable distance (Dcrit) for the data set is given (cf. Figure 3.28). Moderate outliers have DModX-values larger than Dcrit. With process data, moderate outliers often correspond to temporary process upsets, but occasionally more persistent trends or shifts can be diagnosed (Figure 3.29). For process diagnostics, it is of vital importance to uncover outliers in residuals, as a persistently high occurrence of outliers indicates a shift in process behavior.

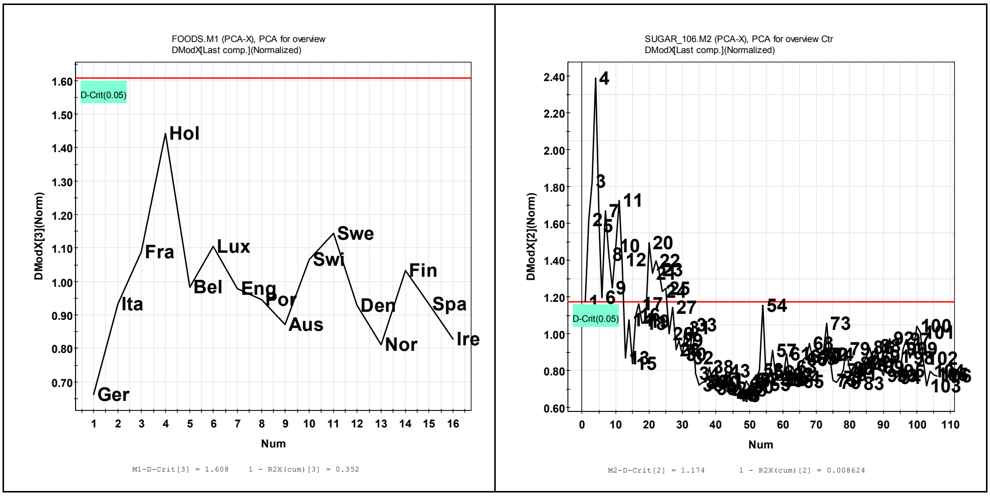


Figure 3.28: (left) DModX control chart for the FOODS data set. No observation exceeds the critical distance (Dcrit). Figure 3.29: (right) DModX chart for a process industry, showing a typical short-term deviation (process upset) at the beginning of the sampling campaign. After a while, the process reaches a state of stability.

Using the SIMCA-P terminology, the calculation of DModX can be formalized as follows. The residual observation variance, S2OX, is computed as

$$S2OX = \sum_k e_{ik}^2 / DF \qquad \text{(eqn. 3.2),}$$

where DF represents the number of degrees of freedom. Here the index *i* represents the observations and the index *k* the variables. The residual observation variance can be converted to the absolute distance DModX as

$$DModX_{abs} = (S2OX)^{1/2} \qquad \text{(eqn. 3.3),}$$

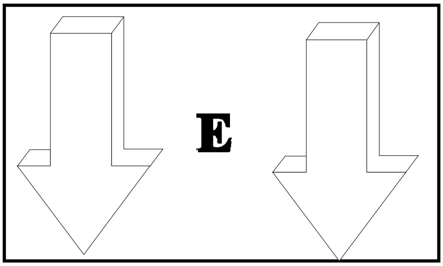
or the normalized distance DModX as

$$DModX_{norm} = [S2OX/variance\ (E)]^{1/2} \qquad \text{(eqn. 3.4).}$$

In SIMCA-P, it is possible to plot S2OX, $DModX_{abs}$, and $DModX_{norm}$ in control charts. A more thorough account of these parameters is provided in Appendix II.

3.5.2 Variable diagnostics – Which variables are well-explained?

Apart from pooling the elements of the E-matrix row-wise, these elements may also be summarized column-wise to produce diagnostics related to the variables (Figure 3.30). One such diagnostic tool is called the *explained variation* of a variable, a quantity which ranges from 0 (no explanation) to 1 (complete explanation). It tells us the extent to which each variable is accounted for by the model.



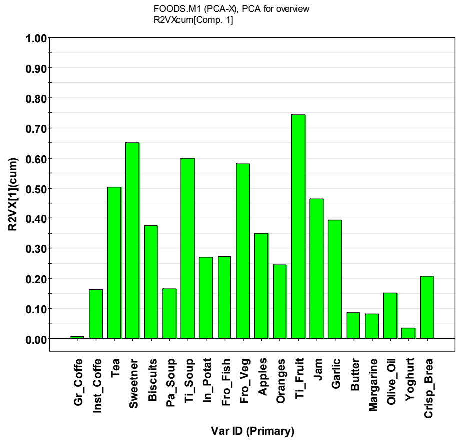
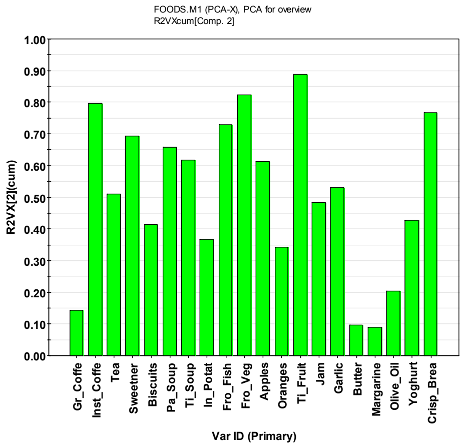


Figure 3.30: (left) In the formation of variable related diagnostics, the entries in matrix E are summarized column-by-column. Figure 3.31: (right) Explained variation of the variables of the FOODS data set after the first PC.

Figures 3.31 – 3.33 demonstrate how the explained variation of each variable in the FOODS data set is altered by increasing the number of principal components in the model. The limiting case is 15 (*N-1*) components because there are 16 (*N*) observations, and in that situation the explained variance of each variable is 1. The latter is an unrealistic model, however.



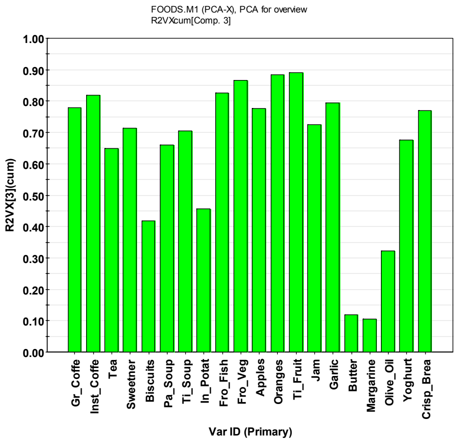


Figure 3.32: (left) Same as Figure 3.31, but after the second component. Figure 3.33: (right) Same as Figure 3.31, but after the third component. Some variables, like Butter and Margarine, are not well explained by the three-component model.

The sequence of Figures 3.31 – 3.33 makes it possible to follow how the individual variables are modelled by the different principal components. Some variables are well accounted for by the first PC, some make their entrance in the model thanks to the second PC, whereas some wait until the last PC. The variable Tin_Fruit is an example of a variable that is predominantly modelled by the first PC, and Gr_Coffe of a variable that is exclusively captured by the third PC. Then, there are variables which load onto more than one component, for instance Yoghurt, which is rather well mapped by the second and the third principal components jointly.

The calculations and the SIMCA-P terminology involved in the derivation of the residuals-based variable diagnostics are given below. By the column-wise summation of the residual elements of **E**, it is possible to describe how well a variable is modelled by the calculation of its explained variation (R^2X) or explained variance (R^2X_{adj}). The explained variance is simply the explained variation adjusted for the degrees of freedom (DF). The values of R_k^2 are related to the loadings. For each component, a, p_{ak}^2 is proportional to how much the kth variable is modelled by this component.

The residual variable variation, *SSVX*, is computed as

$$SSVX_k = \Sigma_i\, e_{ik}^2 \qquad \text{(eqn. 3.5).}$$

This gives the corresponding residual variable variance, $S2VX_k$, by dividing by the degrees of freedom

$$S2VX_k = \Sigma_i\, e_{ik}^2/DF \qquad \text{(eqn. 3.6).}$$

The explained variation and the explained variance of a variable are given by

$$R2VX_k\ (cum) = 1\text{- } SSVX_k[A]/SSVX_k[0] \qquad \text{(eqn. 3.7)}$$

$$R2VXadj_k(cum) = 1\text{- } S2VX_k[A]/S2VX_k[0] \qquad \text{(eqn. 3.8),}$$

where *A* represents the number of principal components. A more thorough description of these parameters is found in Appendix II. Also observe that it is possible to calculate R^2X- and R^2X_{adj}-values pertaining to the complete X-matrix (not just to the individual variables), and this is discussed in Section 3.5.3.

3.5.3 Model diagnostics – How many principal components are really needed?

An important question is *how many components should be included in the model*? This question is linked to the difference between the degree of fit and the predictive ability. The fit tells how well we are able to mathematically reproduce the data of the training set. A quantitative measure of the *goodness of fit* is given by the parameter R^2X (= the explained variation). The problem with the goodness of fit is that with sufficiently many free parameters in the model, R^2X can be made arbitrarily close to the maximal value of one (1.0).

More important than fit, however, is the predictive ability of a model. This can be estimated by how accurately we can predict the X-data, either internally via existing data or externally through the use of an independent validation set of observations. The predictive power of a model is summarized by the *goodness of prediction* parameter Q^2X (= the predicted variation). Here, we use cross-validation (CV) to estimate the predictive ability of the model with increasing number of components (see next section).

The R^2X- and Q^2X-parameters display entirely different behavior as the model complexity increases (Figure 3.34). The goodness of fit, R^2X, varies between 0 and 1, where 1 means a perfectly fitting model and 0 no fit at all. R^2X is inflationary and approaches unity as model complexity (number of model parameters, number of components, …) increases. Hence, it is not sufficient to have a high R^2X. The goodness of prediction, Q^2X, on the other hand, is less inflationary and will not automatically come close to 1 with increasing model complexity. This provided that Q^2X is correctly estimated.

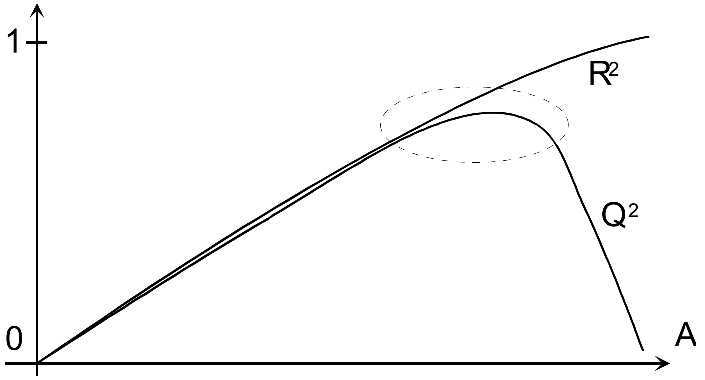


Figure 3.34: The trade-off between the goodness of fit, R^2X, and the goodness of prediction, Q^2X. The vertical axis corresponds to the amount of explained or predicted variation, and the horizontal axis depicts the model complexity (number of terms, number of latent variables, etc). At a certain model complexity, one gets the model with optimal balance between fit and predictive ability.

Hence, by a valid model we mean that it predicts much better than chance. In addition, it should have model parameters with little bias, i.e., they should have the correct sign and be large for important variables and small for unimportant variables. Finally, it should be consistent with fundamental biological, chemical, technical and engineering knowledge.

Using the LOWARP example and a PCA model of the 14 responses, we will now see how the R^2X- and Q^2X-values can be used in the evaluation of a PCA model. Figure 3.35 displays the evolution of the parameters R^2X and Q^2X with increasing model complexity. Clearly, in this case it is not meaningful to use more than three principal components, because the predictive ability does not increase as a result of the fourth and higher-order components.

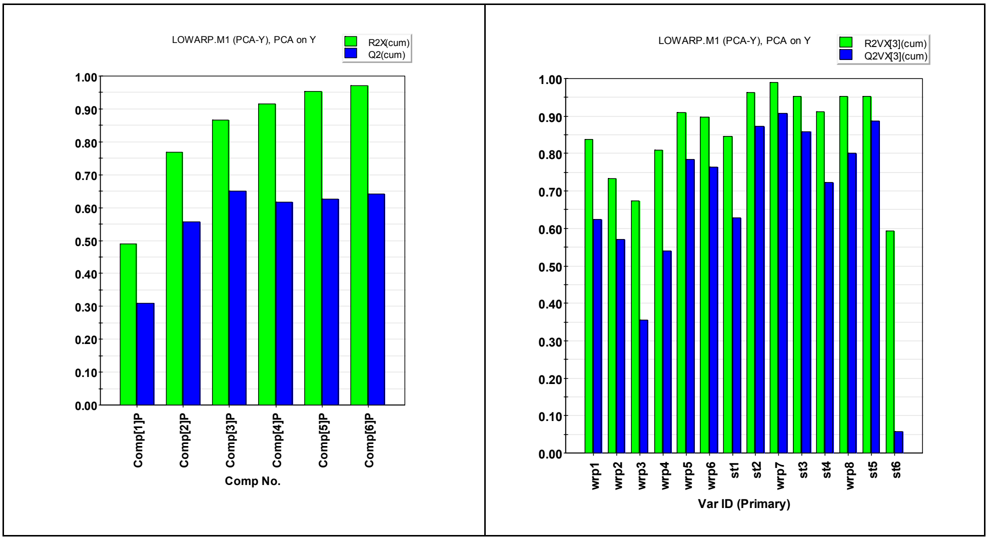


Figure 3.35: (left) Model overview of the PCA model for the 14 responses of the LOWARP data set. The plot shows how the overall R^2X- and Q^2X-statistics change as a function of increasing model complexity. Here, three components appears appropriate, as Q^2X does not increase beyond the third component. Figure 3.36: (right) The overall R^2X- and Q^2X-values may be decomposed into R^2's and Q^2's relating to the individual variables. The situation after three model components is depicted. Apparently, most variables are rather well modelled after three PCs, but there are some exceptions. In particular, the analyst should focus on understanding what is wrong with the variables wrp3, wrp4 and st6, with large gaps between R^2X and Q^2X.

It is possible to partition the overall R^2X- and Q^2X-estimates into values pertaining to the 14 individual variables. Figure 3.36 presents these after three principal components. For the majority of the variables the computed PCA model is able to well describe their variation. But there are some exceptions. In particular, the three variables wrp3, wrp4 and st6 have large gaps between R^2X and Q^2X. Such large gaps indicate problems with these variables.

3.5.4 Cross-validation

The approach to finding the optimal model dimensionality advocated throughout this course book is called cross-validation [Wold, 1978]. Cross-validation (CV) is a practical and reliable way to test the significance of a PCA- or a PLS model. This procedure has become standard in multivariate analysis, and is incorporated in one form or another in most commercial software. However, CV is implemented differently in different packages, which may cause some confusion when comparing models developed by different packages.

With CV the basic idea is to keep a portion of the data out of the model development, develop a number of parallel models from the reduced data, predict the omitted data by the different models, and finally compare the predicted values with the actual ones (Figure 3.37). The squared differences between predicted and observed values are summed to form the *predictive residual sum of squares* (PRESS), which is a measure of the predictive power of the tested model. PRESS is computed as

$$PRESS = \Sigma (x_{ik} - \hat{x}_{ik})^2 \qquad \text{(eqn. 3.9).}$$

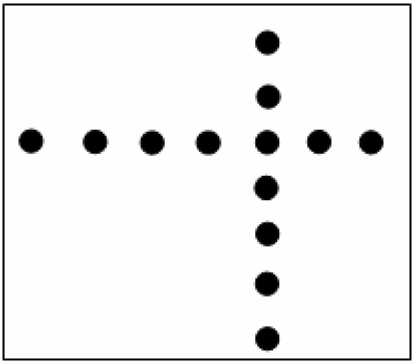


Figure 3.37: In PCA, data are divided into G groups, typically between 5 and 10 groups, and temporarily deleted in a combined row-wise/column-wise fashion. A model is established for the data devoid of one group. The deleted group of elements is predicted by the model and the predictive residual sum of squares (PRESS) is calculated. This procedure is repeated G times, followed by the summation of all partial PRESS-values in terms of an overall PRESS-value. If a new PC_a enhances the predictive power compared with the preceding PC_{a-1}, the new PC_a is kept in the model.

In SIMCA-P, CV is conducted for each consecutive model dimension starting with A = 0. For each additional dimension, CV gives a PRESS, which is compared with the residual sum of squares (RSS) of the previous dimension. When PRESS is not significantly smaller than RSS, the tested dimension is considered insignificant and the model building is stopped.

Normally, the performance a PCA model in SIMCA-P is evaluated by simultaneously considering the explained variation R^2X (goodness of fit) and the predicted variation Q^2X (goodness of prediction). As shown by equations 3.10 and 3.11, these two statistics resemble each other:

$$R^2X = 1 - RSS/SSX_{tot.corr.} \qquad \text{(eqn. 3.10)}$$

$$Q^2X = 1 - PRESS/SSX_{tot.corr.} \qquad \text{(eqn. 3.11),}$$

and they are both dimensionless. In the expressions above, $SSX_{tot.corr.}$ represents the total variation in the X-matrix after mean-centering.

In the evaluation of the parameters R^2X and Q^2X, there are a few noteworthy facts. The first is that without a high R^2X it is impossible to get a high Q^2X. Generally, a $Q^2X > 0.5$ is regarded as good and a $Q^2X > 0.9$ as excellent, but these guidelines are of course heavily application dependent. Finally, the difference between R^2X and Q^2X must not be too large, and preferably not exceeding 0.2 - 0.3.

3.6 Questions for Chapter 3

1. Which two approaches are commonly used for data pre-processing?
2. What is a principal component?
3. What is a *score*?
4. What is a *loading*?
5. What is a strong outlier?
6. Which detection tool can be used to uncover strong outliers?
7. What is a moderate outlier?
8. Which detection tool can be used to identify moderate outliers?
9. Which statistics indicate which variables are well explained?
10. How do we determine the optimal dimensionality of a PCA model?

3.7 Summary and discussion

Principal component analysis summarizes the variation of a data matrix X, as a product of two low-dimensional matrices, T and P´, which may be easily overviewed and used. The data in the multi-dimensional space, defined by the measured variables, is modelled as a plane or hyperplane, the axes of which are called the principal components. Each principal component can be displayed graphically and may often be interpreted according to chemical, technical and/or biological knowledge.

Prior to PCA, data are typically pre-processed by means of mean-centering and scaling to unit variance. With PCA the X-matrix is then decomposed ($X = 1*\overline{x}' + T*P' + E$) into the product of two matrices, the (N x A) score matrix T times the (A x K) loading matrix P´, plus an (N x K) "noise" matrix of residuals, E. The value of A, the number of principal components, is usually determined by cross-validation. As long as the dimensionality of the projection, A, is small, this model of the variation in X as T*P´ is a considerable simplification.

Plots of PCA scores show the relationships among the observations including outliers. Plots of residuals, for example DModX, pinpoint moderate outliers. To support decision making some statistical tools are available. For score plots, Hotelling's T^2 may be used to define a 95% or 99% tolerance region. For DModX, a critical distance corresponding to the 0.05 level is available. Observations outside the ellipse defined by Hotelling's T^2 deviate from normality, and observations exceeding the critical distance in DModX do not fit the model well.

We make a distinction between strong and moderate outliers. Strong outliers, seen in score plots, *conform* with the overall correlation structure of the data (although they may have an extreme character), whereas moderate ones, which are found in residual plots, *break* this correlation structure. A moderate outlier does not have the same profound effect on the PCA model as a strong outlier does. But for diagnostic purposes it is important to identify outliers in residuals, because they indicate lack of homogeneity ("upsets") in X.

Another attractive property of PCA is that it applies to almost any type of data matrix, e.g., matrices with many variables (columns), many observations (rows), or both. The precision and reliability of the PCA parameters related to the observations (scores, DModX) is enhanced by increasing the number of relevant variables. This property is readily understood by realizing that the "new variables", the scores t_a, are estimated as weighted averages of the X-variables. Any (weighted) average becomes more precise the more numerical values are used as its basis. Hence, PCA works well with "short and wide" matrices, i.e., matrices with many more columns than rows.

Analogously, the PCA parameters supplying information related to the variables, for example, loadings, R^2X, Q^2X, etc., become more reliable the larger the number of observations. This because the loadings, etc., are linear combinations, i.e., *weighted averages*, of the *N* observation vectors. Thus, PCA also works well with "long and lean" data structures, i.e., data tables with many more rows than columns.

Obviously, PCA is also capable of handling "square" matrices, i.e., situations in which the number of rows and columns are approximately equal.

Because of its power and versatility PCA today goes under many names. Singular value decomposition (SVD) is a common name in numerical analysis, Karhunen-Loeve expansion in electrical engineering, and Eigenvector analysis is often used in the physical sciences. In some French-speaking countries a special double-scaled variant of PCA, known as Correspondence analysis, is much favored (see Chapter 10). And, finally, we have the closely related method Factor analysis (FA), which was originated by psychologists in the 1930s. PCA and FA are often confused and often the two names are incorrectly used interchangeably [Jackson, 1991].

PCA also forms the basis for hierarchical multivariate modelling (see Chapter 24). With the hierarchical approach score variables, derived from multiple PCA models at the lower modelling level, are concatenated to form new upper modelling level data matrices that are again analyzed by PCA. The upper level provides a superficial overview of the relationships among the different blocks of variables. The lower level allows zooming-in onto the most influential variables in the important variable blocks. The use of PCA as a corner-stone in data mining and data integration is delineated in Chapter 27.