MULTIVARIATE DATA ANALYSIS WITH PCA, CA AND MS

TORSTEN MADSEN ©2007

Archaeological material that we wish to analyse through formalised methods has to be described prior to analysis in a standardised, formalised way. We describe *units* in terms of *variables*. The units may be proper physical objects like pots, swords, axes, brooches, etc. or it may be assemblages of physical objects like graves, hoards, pits, or indeed any kind of excavation contexts. With reference to a modern object-oriented approach to the world we simply term all these kinds of units as *objects*. The variables can be any abstracted quality from the units like measurements, discrete descriptive elements or types of objects.

Our formalised description is an abstraction from the archaeological material background. We analyse this abstraction in order to isolate meaningful structure that will provide us with means to understand and interpret the material background. There are many ways in which we may do formalised analyses of our description. Some are uni-variate in the sense that they focus on one variable at a time. Others are bi-variate, analysing variables pair wise. A third group is multivariate, analysing three or more variables together. This paper deals with three multivariate methods – Principal Components Analysis (PCA), Correspondence Analysis (CA) and Metric Scaling (MS). Apart from introducing and discussing the three types, a number of examples will be presented. The data for these can be found in the file *Examples.xls*.

The central computational model for all three methods is the same. It is based on what is called a singular value decomposition of a matrix. The differences between the methods are based on the pre-treatment of the input data, as each method is aimed at data with particular statistical qualities.

Principal components analysis

The PCA method is designed to isolate patterns of covariance in a set of measurement variables, i.e. we expect the state of variables to be dependent on each other in some way. The state of one variable for an object has implications for the state of other variables for the same object. Our data could be a number of pots described by rim diameter, base diameter, shoulder diameter, total heights, etc. If we know the rim diameter of a pot we also know something about most other measurements of the pot. As a minimum, and quite trivial, the effect of general size will make most if not all measurements of big pots larger than those of small pots, but beyond size we may find more interesting patterns of covariance between the measurements.

Crucial to PCA is the basis on which the variables are compared. This is determined trough a matrix of coefficients expressing the degree of covariance pair wise between all variables. There are two types of coefficients that may be used here. One is the covariance coefficient, and the other the correlation coefficient (Persons r). The result will normally differ between the two, and it is therefore necessary to understand the difference between them in order to decide when to choose the one or the other.

Correspondence analysis

Another typical set of data in archaeology is objects described by counts (including presence/absence) of some characteristic elements of the objects. It could be graves described in terms of their content of artefact types like different types of hair pins, brooches, belt buckles, weaponry, pottery, everyday utensils, etc. If in a grave, we find a typical female ornament like a hairpin it is quite likely that the same grave may contain another female ornament like a brooch, but quite unlikely that it will also contain a sword. In contrast if a grave contains a sword it is highly unlikely that it will also contain a hairpin or a brooch, but quite likely that it will contain a belt buckle. Covariance in type inventories due to sex is one of the most common structuring elements in graves. But there are certainly others like social ranking, and indeed chronological changes. Correspondence Analysis (CA) is a method designed to isolate patterns of covariance in a table of incident variables (presence/absence or counts) recorded for a number of objects. Such tables are also known as *contingency* tables.

In order to investigate this type of data, correlation coefficients as used in PCA is of no use. The data are far from normally distributed. On the contrary they are heavily skewed to the left (i.e. towards small values) more or less following the Chi-Square distribution, and the Chi-Square statistic, ideally suited to deal with contingency tables, is in fact used in CA. The staring point of a CA is a table where each cell is computed as:

$$y_{ij} = \frac{(O_{ij} - E_{ij})}{\sqrt{E_{ij}}}$$

For each cell the observed value minus the expected value is divided by the squarer root of the expected value. Expected values are derived from the row and column sums of the table of observed values under the assumption of a random non-structured distribution of values across rows and columns.

Metric scaling

Scaling methods use similarity coefficients (or distance coefficients) as their starting point. A similarity coefficient is a numerical value that expresses the similarity between two objects. Mostly, similarity coefficients are structured in such a way that they attain the value of 1 if the objects are identical and the value of 0 if the objects have nothing in common. (Distance coefficients are in principle merely reciprocals of similarity coefficients).

The advantage of similarity coefficients is that they can easily be constructed in such a way that information from variables on different types of scale can be combined. Thus it is possible to analyse measurement data together with counts from contingency tables. However, and this is a big disadvantage, when constructing the similarity coefficients the connection between objects and variables is broken. A coefficient is a general expression of similarity between two objects calculated from the state of their variables. Afterwards it is not possible to see the contribution of the individual variables, and these are completely left out of the analysis.

The computational background to PCA, CA and MS

It is quite difficult for a non mathematician to grasp the rationale behind the methods let alone the actual computations. The following is an attempt to make a very informal introduction to the three methods. The core of computation for all three is identical, but for clearness of presentation the following will be worded along the lines of a PCA. Later, when the three methods are exemplified, I will go into more detail with the characteristics of the individual methods and the differences in output they produce. If you wish to get a more appropriate introduction to these methods M.J. Baxter’s Exploratory Multivariate Analysis in Archaeology (1994) can be recommended.

If you have two variables like *Rim diameter* and *Neck diameter* describing a series pots (objects) you may depict their interrelationship in a two dimensional plot with each set of linked observations of *Rim diameter* and *Neck diameter* shown as points. You will probably find that the points tend to form a linear configuration due to covariance between the two variables. This linear trend can be described with a line known as a regression line through the point scatter based on some criterion of fit. This line can be seen as a one dimensional representation of information stemming from two dimensions. The representation will not be perfect of course. The points will be scattered to both sides of the line with varying distances known as residuals.

A traditional regression line takes its starting point in one of the variables (the independent variable, by convention placed on the x axis) and use a criteria of fit based on the dependent variable (y axis) exclusively (fit is measured as distance from points to regression line parallel to the axis of the dependent variable). Thus the result will differ according to which variable is chosen as the independent.

There is another regression method called orthogonal regression, where the variables are independent. Finding this regression line is based on a criterion of fit where the distance from the points to the regression line has to be measured perpendicular to this. The criterion is that the squared sum of distances from the points to the regression line is a minimum. It is obvious that finding the solution to this problem is not trivial as the criteria of fit is measured perpendicular to the line we seek, and it is hence impossible to set up a simple formula, because we do not know in which direction to measure. It can be shown, however, that the solution will be one of a simple rotation of the axes describing the two variables if their point of origin is shifted to the common centre of mean values. What is obtained are two new axes perpendicular to each other (exactly as the originals), where the first covers the maximum part of variation in the point scatter and the second the residuals.

It is not difficult to imagine that the principle of orthogonal regression will work with three variables depicted in a three dimensional space as well. Following the orthogonal regression – rotation of the axes around the centre of mean values – the leading axis in the rotation will cover the maximum part of the variation of the three original variables. The second axis will cover the maximum part of the remaining variation, and the third axis the rest. In the process of orthogonal regression we aim to represent as much information as possible on the first axis, as much of the remainder information as possible on the next axis, etc. We call the leading axis in the rotation for the first *Principal axis*, the next for the second *Principal axis*, etc.

Obviously, we are not satisfied with analysing just three variables together. However, bringing more than three variables into an analysis blocks our visual geometric understanding. We have to look at the problem arithmetically, which also of course is the way we have to deal with it computationally.

If we go back to the two dimensional case it is fairly easy to see that the two new principal axes (call them P_1 and P_2) being a transform (rotation) of *Rim diameter* and *Neck diameter* must relate to these original variables in a unique way that can be described through simple linear equations. Thus the new axes or components as they are called, when we view the problem arithmetically, P_1 and P_2 are constituted by linear combinations of *Rim diameter* and *Neck diameter*:

$P_1 = a_1$ *Rim diameter* + a_2 *Neck diameter*
$P_2 = b_1$ *Rim diameter* + b_2 *Neck diameter*

where a_1, a_2, b_1 and b_2 are positive or negative values.

Now if we have three variables like *Rim diameter, Neck diameter* and *Shoulder diameter* we will just have to add a new element to the equations and at the same time of course we get three principal components:

$P_1 = a_1$ *Rim diameter* + a_2 *Neck diameter* + a_3 *Shoulder diameter*
$P_2 = b_1$ *Rim diameter* + b_2 *Neck diameter* + b_3 *Shoulder diameter*
$P_3 = c_1$ *Rim diameter* + c_2 *Neck diameter* + c_3 *Shoulder diameter*

When we view the problem in this manner there is obviously no limit to the number of variables we can include. We simply add a new principal component and a new element to each of the component equations.

Still we have to find the values of a_1, a_2, a_3 etc. and the problem does not become less by turning it from a geometrical problem into an arithmetical problem. The solution lies within matrix algebra and specifically with a unique factorisation of a matrix called *singular value decomposition*. The singular value decomposition of a matrix (table of input data in our case) can be shown to produce the components we look for. For an introduction to matrix algebra and singular value decomposition you should read Baxter 1994, Appendix B.

There are different algorithms for performing singular value decomposition in a computer. The most economically of these in terms of speed and storage uses iterative procedures that gradually converge towards the desired result, and stop when a control value shows that the result obtained is satisfactory. Although generally very stable and reliable, it occasionally happens that convergence is not reached, and calculations have to stop without result. The algorithm used in CAPCA is one published by Wright in 1985.

To gain further insight into this type of multivariate analysis and the kind of information it provides I will take the example with the pot diameters a little further. In doing so I will move directly into the realms of PCA, and part of what is said here will be repeated when we turn to the actual examples of this specific type of analysis.

The new set of axes created by singular value decomposition is technically referred to as e*igenvectors,* but in connection with a PCA they are called *principal components*. The principal components are ranked in such a way that the first component covers the largest part of the total variation in the data set, the second component the second largest part, etc. They may be viewed as a new set of variables substituting the original ones, and in doing so they retain the total amount of variability in the data, but represent it in a different, more structured way.

The values in front of each of the original variables in the equations of the principal components are called *loadings*. One of the things they show is how large a part of the variation of the original variables is represented in the new principal components. To see how, you should “read” vertically for each of the original variables. With reference to the equations below, for each of the original variables the sum of squared values will amount to 1 (= 100%) and the percentages of variation in *Rim diameter* that goes into P2 is thus the square of -0.26, which equals 0.07 (= 7%)

P1 = 0,96 *Rim diameter* + 0,95 *Neck diameter* + 0,97 *Shoulder diameter* + 0,88 *Height*
P2 = - 0,26 *Rim diameter* - 0,28 *Neck diameter* - 0,22 *Shoulder diameter* + 0,32 *Height*
P3 = 0,11 *Rim diameter* + 0,02 *Neck diameter* + 0,00 *Shoulder diameter* - 0,34 *Height*
P4 = - 0,05 *Rim diameter* - 0,03 *Neck diameter* + 0,11 *Shoulder diameter* - 0,02 *Height*

Loadings have much the same qualities as correlation coefficients. Not only do they tell by their size (between 0 and 1) how strong the correlation is, but also by their sign whether it is a positive (when one grows the other grows as well) or negative (when one grows the other diminish) correlation. In the above example we find that all variables have a strong positive correlation with the first principal component (heights a little less than the diameters). For the second principal component on the other hand there is a weak negative correlation with the diameters and a weak positive correlation with heights. To help understand how the original variables are structured in relation to the principal components it is often a help to view the loadings in two way plots.
If we take the sum of squared values for each principal component we get what is termed the *eigenvalue* of the component (also occasionally referred to as *latent root*). This can tell us how large

a part of the total variation a principal component represents. As the total variation of the example above is 4 (1 for each of the original variables) and the eigenvalue of P1 is 3.54 (sum of squared values) then P1 is accounting for 88% of the total variation.

The values of the objects – here the individual pots – on the principal components are called *Scores*. They are calculated by substituting the variables in the equations with the actual values for the individual objects. The scores will consist of a blend of positive and negative values and will bear no resemblance to the input values. The main reason for this is that as a minimum all input values has been centred (by subtraction of their mean value) and most likely also standardised (by division with their variance or standard deviation). Otherwise they have not been altered. If you could make the mind experiment of plotting the pots in a four dimensional space of the original variables and then see them in the four dimensional space of the new principal components you would find that they would display the exact same spatial structure. Only the axes would be directed differently.

As with the original variables, a meaningful way to view data is through two way plots of the principal components. The major difference, however, is that whereas you have to plot all original variables against one another to gain an overview, you only have to plot the first few components, and possibly only the first two to view the structure of the data. In the above example the two first principal components covers as much as 96% of the total variation indicating that the two last components are of no interest at all. The analysis has thus very effectively reduced the number of dimensions needed to give an adequate representation of the information. This capability of representing the important part of the variation in complex data sets by way of a few new principal components is the hallmark of this type of multivariate method. It makes it a very efficient tool to seek structure in data.

PRINCIPAL COMPONENT ANALYSIS

The optimal type of data for PCA is measurements of some kind. Other kinds of quantitative data can also be analysed including indexes and counts, but the latter type of data is far more suited for CA and should preferably be analysed through this. PCA is a classic, and as such it has been used intensively, and consequently often with little regard to the nature of the data analysed. It will probably help here if it is understood what happens to the data you input prior to analysis in PCA. If we return to what was said earlier about finding the orthogonal regression lines one change to the data is a must. For each variable we have to subtract the mean value from each value of the variable in order to centre the variable on its mean. Thereby we create a common centre through which all variable axes pass. The actual formula used is:

$$X_i = \frac{(X_i - \overline{X})}{\sqrt{n-1}}$$

The division by the square root of the number of instances is a division by a constant, since it will be the same for all variables. It thus does not change the overall structure. If we exclusively use this kind of transformation to the data, the PCA will be based on what is known as a covariance matrix.

We may, however, also choose to standardise data. Standardisation means that all variables apart from being centred also have unity dispersion. That is they have all a standard deviation of 1. To obtain this we use the following formula:

$$X_i = \frac{(X_i - \overline{X})}{\sqrt{\frac{\sum_1^n (X_i - \overline{X})^2}{n-1}}\sqrt{(n-1)}}$$

Using this standardisation we radically change the absolute size of values. They now are all more or less equal in weight, exactly as if we had done percentage calculations on counts. If we input data standardised in this way, the PCA will be based on what is known as a correlation matrix.

The standard deviation and consequently the correlation coefficient can be considered correct only if the variables are reasonably normally distributed. Therefore we have to check the measures of skewness and kurtosis, both of which have to be not two far from zero. Any measure between 1 and -1 are fine for our purpose and even a few positive or negative readings of 2 or 3 should not disturb us. However, if the absolute values become very high or if there are many between 1 and 3 we should consider not using the correlation matrix. Before doing so we should check, however, if scale transformations of the variables can solve the problem. In CAPCA you can turn on automatic scale transformation, and using Log10, Ln or ArcSin transformations the skewness and kurtosis is normally reduced to acceptable levels.

For a PCA, weighting of variables beyond that of standardisation should normally not be considered. Weighting of objects on the other hand is quite feasible and legitimate. A good case is seen below in the example of Neolithic pots.

It is standard procedure to plot two variables against each other in order to evaluate the value distribution of the objects with respect to the two variables. It is obvious that we can do the same with principal components. Normally we will only make a plot of the first and second principal component against each other, and possibly the second against the third to see if interpretable information should exits on the latter. The plots can be looked upon exactly as plots of the original variables, but you won’t find any likeness between the values of the component axes and the values of the original variables. Neither do the sign of the values (negative or positive) mean anything by itself. In fact you will often see sign reversals in the output after even minor changes to the input (se the example on Neolithic pots below).

The original variables cannot be plotted together with the objects (as it is possible with CA). You can however create a so called *biplot* of these variables as vectors in the n dimensional space created by the principal components. Normally you will only inspect the plot of variables against the two first principal components. Each variable will be represented by a point, but you should imagine lines (vectors) reaching from 0,0 in the plot to the points. This vector plot should be interpreted in terms of correlation/covariance. Vectors in the same direction has positive correlation/covariance, Vectors in opposite directions have negative correlation/covariance. Vectors perpendicular to each other have zero correlation/covariance. Long vectors have strong positive or negative correlation/covariance. Short vectors have small positive or negative correlation/covariance.

The only connection between the objects and variable plots lies in the orientation defined by the principal components. Thus a variable vector stretching along, say the positive part of the first principal component, indicate that the objects lying in the same direction in the objects plot will have high values for this variable, while those lying in the opposite direction will have low values.

Example using measures from 430 Iron Age lance heads

To explore the difference of using PCA based on a covariance matrix and a correlation matrix, and to have a closer look at the information we receive from PCA we will look at an example with spearheads from the Iron Age votive bog finds from Illerup, Eastern Jutland, Denmark (Ilkjær

1990). The 430 spearheads are described by nine different measures (Ilkjær 1990: 30), and they all belong to three specific types (14, 15 and 18) defined in the publication based on selected criteria's among these measures. Thus a PCA of all the measurement data ought to produce a result in accordance with the type division.

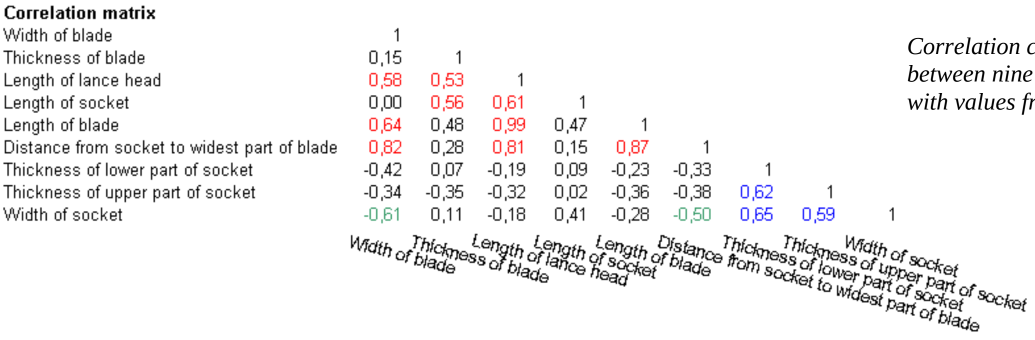
We start out with a PCA based on a correlation matrix. As the variables become standardised through the standard deviation they should be as close to a normal distribution as possible. The function of automatic transformation in CAPCA has therefore been turned on. It investigates if a Log10 transformation, a Ln (natural logarithm) transformation or an Arc Sin transformation will provide better normality than untransformed data, and if so it will substitute the original data with the best transformation. Since all data are standardised such a transformation will not change the structure of the data.

	Skewness	Kurtosis	
Width of blade	-0,36	0,31	No transformation
Thickness of blade	-0,39	4,23	No transformation
Length of lance head	-0,04	0,65	No transformation
Length of socket	-0,01	1,98	No transformation
Length of blade	0,20	0,43	No transformation
Distance from socket to widest part of blade	0,09	-0,47	No transformation
Thicknes of lower part of socket	-2,02	24,03	ArcSin transformed
Thickness of upper part of socket	-0,99	13,80	ArcSin transformed
Width of socket	0,32	5,53	ArcSin transformed

Skewness and kurtosis for values of nine measurement variables with values from 430 spearheads.

Looking at Skewness and Kurtosis the main measurements – *Width of blade*, *Length of lance head*, *Length of socket*, *Length of blade* and *Distance from socket to widest part of blade* – are well behaved. The measurements *Thickness of blade*, *Thickness of lower part of socket*, *Thickness of upper part of socket* and *Width of socket*, which all have a very narrow measuring range, all have a high positive kurtosis indicating that their distributions are too high and narrow for normality. Three of them have been ArcSin transformed, but with little result.

The correlation coefficient matrix is significant for understanding the result of PCA. It gives an immediate impression of which variables co-vary and whether the association is positive or negative. One way of using this matrix is to outline clusters of high positive or negative values by adding colours. Colouring important coefficients is an efficient way to outline the structure of co-variation among the variables. In general the method does not work with a covariance matrix, because the size of the coefficients here varies with the absolute measuring range of the variables and because most variables tend to be positively correlated no matter what variation may be uncovered.



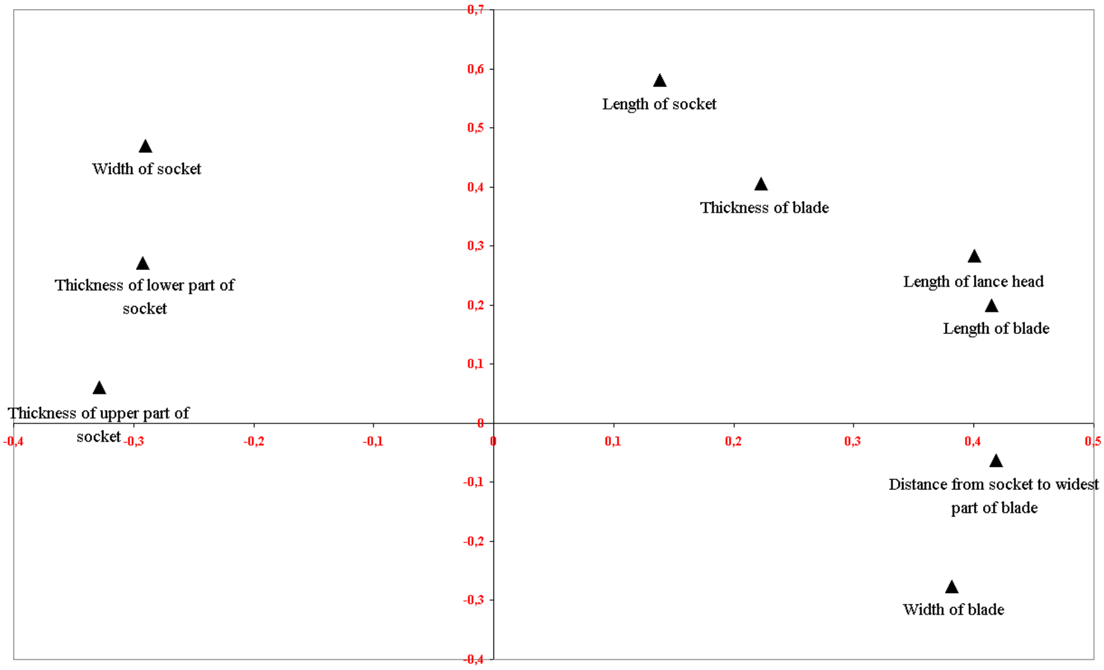
Correlation matrix	Width of blade	Thickness of blade	Length of lance head	Length of socket	Length of blade	Distance from socket to widest part of blade	Thickness of lower part of socket	Thickness of upper part of socket	Width of socket
Width of blade	1								
Thickness of blade	0,15	1							
Length of lance head	0,58	0,53	1						
Length of socket	0,00	0,56	0,61	1					
Length of blade	0,64	0,48	0,99	0,47	1				
Distance from socket to widest part of blade	0,82	0,28	0,81	0,15	0,87	1			
Thickness of lower part of socket	-0,42	0,07	-0,19	0,09	-0,23	-0,33	1		
Thickness of upper part of socket	-0,34	-0,35	-0,32	0,02	-0,36	-0,38	0,62	1	
Width of socket	-0,61	0,11	-0,18	0,41	-0,28	-0,50	0,65	0,59	1

Correlation coefficient matrix between nine measurement variables with values from 430 spearheads.

Looking at the correlation matrix we find two distinct clusters of coefficients. The major (red) link together the main measurements with positive correlations. Thus there are high correlations between *Length of lance head*, *Length* of blade and *Distance from socket to widest part of blade*. Linked to this cluster are also *Length of Socket* and *Thickness of blade*. Another cluster (blue) with internal positive correlation consists of *Thickness of lower part of socket*, *Thickness of upper part of socket* and *Width of socket*. There is a negative correlation between this cluster and the major one most clearly expressed by the correlation (green) between *Width of socket* on the one side and *Width of blade* and *Distance from socket to widest part of blade* on the other. Thus there seems to be a tendency that plump sockets go with slender and partly smaller blades and *vice versa*.

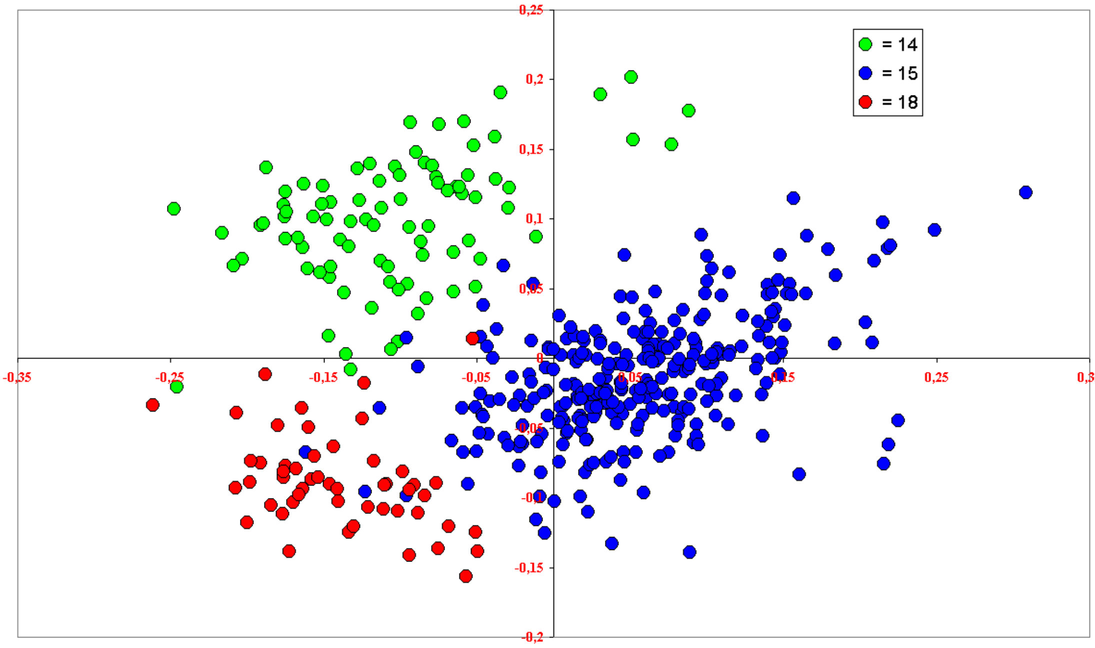
	1. Axis	2. Axis	3. Axis	4. Axis
EigenValues	4,67	2,10	0,99	0,60
Explanation %	51,85	23,30	10,95	6,64
Cumulative Explanation %	51,85	75,15	86,10	92,74

Four principal components have been calculated, and looking at the Eigenvalues this appears to be enough to represent all important information. In fact the first two components cover 75% of all information, and it should be sufficient to study a graphical representation of these two axes to evaluate the result of the analysis.



Biplot of variable loadings from a PCA based on correlation coefficients. Data consist of 430 spearheads measured by nine variables.

What can be inferred from the correlation matrix is clearly displayed in the variable plot. The main cluster of strongly correlated measures is seen to the right, and the cluster of socket thickness and width measures is seen to the left. The negative correlation between the two clusters is shown by their position on each side of 0 on the first principal component. On the second principal component there is a positive correlation between *Width of socket*, *Length of Socket* and *Thickness of blade* suggesting that these tend to vary together. In opposition lies *Width of blade* indicating that wide blades are generally thin and goes with short sockets.



Objects plot of a PCA based on correlation coefficients. Data consist of 430 spearheads measured by nine variables.

Looking at the objects plot we find that the three types are fairly well separated through the analysis. They were originally defined by setting hierarchical ordered thresholds for selected measures (Ilkjær 1990 p. 42, Abb. 28). The classification scheme is very complex and especially for type 14 difficult to follow. No less than four different sets of hierarchically organised criteria can lead to the type. Using the objects plot together with the variables plot we can provide a general characteristic of the three types. Type 14 tends to have relatively short, narrow but thick blades and long and plump sockets. Type 15 tends to have long wide blades with a fairly slender socket. Type 18 tends to have short, relatively wide blades and a short plump socket.

We will now rerun the analysis based on the covariance matrix instead of the correlation matrix.

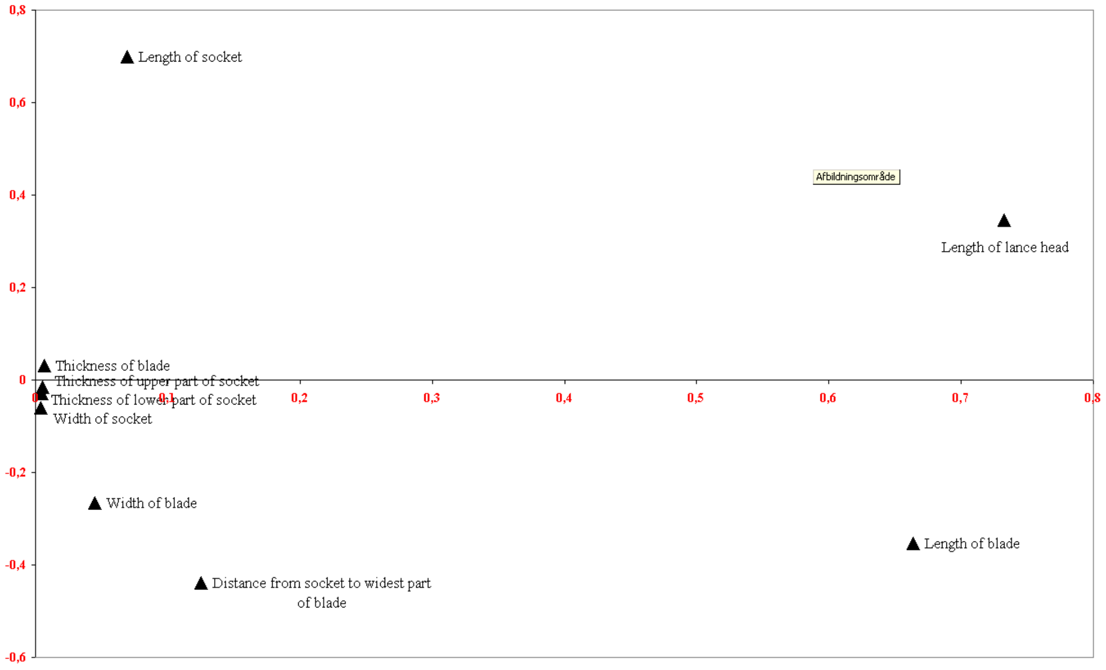
Covariance matrix	Width of blade	Thickness of blade	Length of lance head	Length of socket	Length of blade	Distance from socket to widest part of blade	Thickness of lower part of socket	Thickness of upper part of socket	Width of socket
Width of blade	78,33								
Thickness of blade	2,35	3,10							
Length of lance head	442,56	80,47	7461,49						
Length of socket	0,46	15,09	808,97	238,20					
Length of blade	442,07	65,38	6652,08	570,88	6080,67				
Distance from socket to widest part of blade	126,16	8,65	1224,11	41,49	1182,54	303,41			
Thickness of lower part of socket	11,51	0,80	61,02	3,86	57,14	16,23	4,20		
Thickness of upper part of socket	6,85	1,65	60,07	4,16	55,90	12,59	1,86	2,22	
Width of socket	11,70	0,36	44,10	-6,03	50,12	17,35	2,48	1,57	3,05

Covariance coefficient matrix between nine measurement variables with values from 430 spearheads.

Looking at the covariance matrix we can immediately see that it differs a lot from the correlation matrix. In the diagonal cells, where, due to standardisation, the correlation matrix held 1's, we find numbers that are an expression of the value range of the variables (actually the average squared distance of the individual values from the mean value of the variable). We can see that the size effect also affects the coefficients of the off diagonal cells. Thus cells combining *Length of Blade* and *Length of lance head* with other variables have considerably higher coefficients than any other cells. A coefficient, however, is not merely a reflection of the combined value ranges of two variables, as can easily bee seen from some of the smaller coefficients. We are dealing with a measure of covariance. Since no standardisation has taken place, however, the range of values that the individual variables cover does influence the result of the analysis. If this turns out to be a problem you should use correlation coefficients.

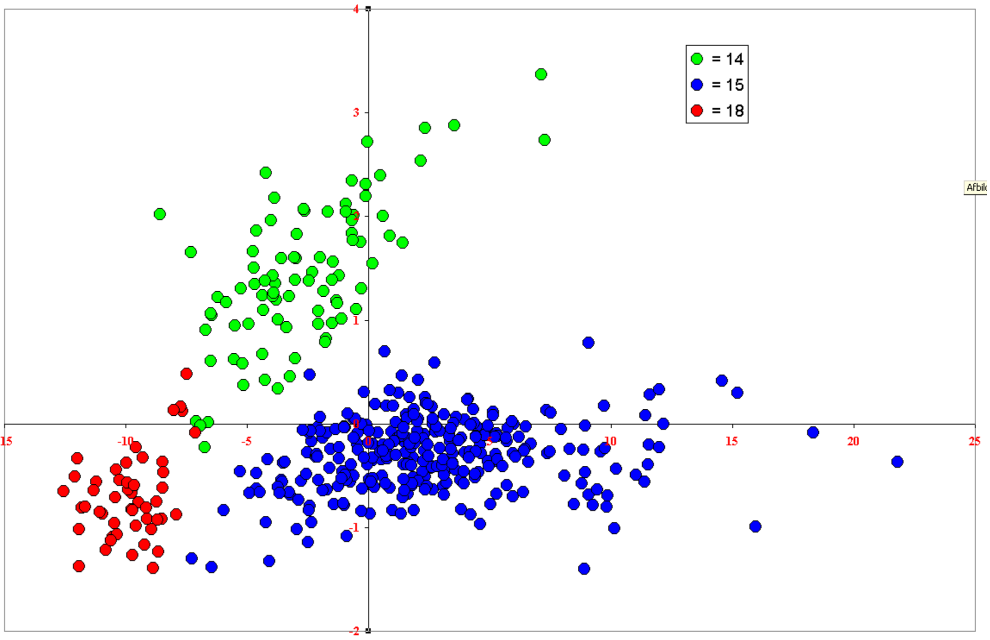
	1. Axis	2. Axis	3. Axis	4. Axis
EigenValues	13498,33	316,19	49,58	16,98
Explanation %	97,20	2,28	0,36	0,12
Cumulative Explanation %	97,20	99,48	99,84	99,96

If we look at the eigenvalues it is obvious that the size effect of value ranges has ended up on the first principal component. An explanation % of 97 compared to 52 with the analysis based on correlation coefficients says it all. In this case we cannot use the eigenvalues to decide how many components we need to cover the important information. As we shall see the second component holds a lot of important information.



Biplot of variable loadings from a PCA based on covariance coefficients. Data consist of 430 spearheads measured by nine variables.

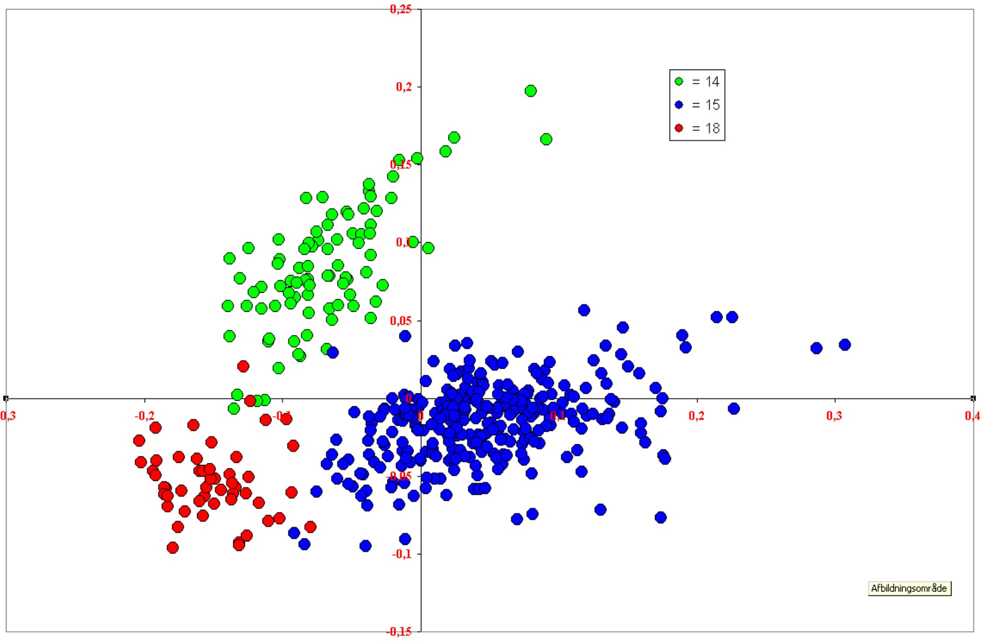
The first thing to note is that the two marked clusters of variables we found in the first analysis at each end of the first component are gone. Instead the first component presents a fairly direct reflection of the size of the value ranges of variables with the two really big ones far to the right. In accordance with what usually happens, when size plays a role, all variables are positively correlated. As for the second principal component we find *Length of socket* negatively correlated with *Width of blade* and *Distance from socket to widest part of blade* exactly as in the first analysis, but in this case *Length of lance head* and *Length of blade* follow this split as well.



Objects plot of a PCA based on covariance coefficients. Data consist of 430 spearheads measured by nine variables.

The surprise comes with the plot of objects. It is a far better result than the one obtained with the correlation matrix. Not only are the groupings more distinct, but there are no longer apparent “misclassifications” (e.g. objects classified as type 15 lying among objects of type 18). Why is this so? One obvious explanation is that the standardisation taking place in the first analysis gives too much influence to unimportant variables like the various sickness measures. Incidentally, these are also the variables that are far from normality, and thus not suited for a PCA based on correlation coefficients. To test this I have run two new analyses where *Thickness of lower part of socket*, *Thickness of upper part of socket*, *Thickness of blade* and *Width of socket* have been excluded.

The analysis using the covariance matrix shows no changes at all in the objects plot and the only change in the variables plot is that the four excluded variables have disappeared. The analysis using the correlation matrix, however, changes a lot. The plot of the objects looks as follows:



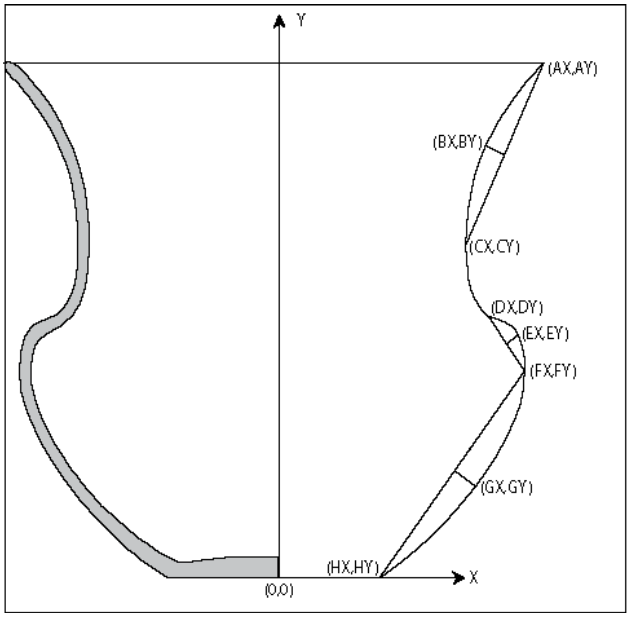
Objects plot of a PCA based on correlation coefficients. Data consist of 430 spearheads measured by five variables.

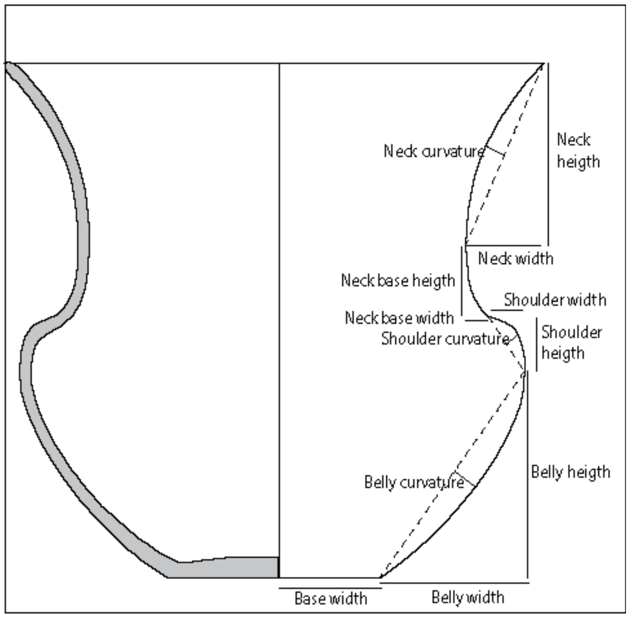
The separation of groups is now very good and there is hardly any overlap between types. It looks very much like the plot from the analysis of the covariance matrix, but it is clearly not identical. It is difficult to say which we should prefer.

This example clearly demonstrates the differences and problems with the two input types. Correlation matrix input is the “cleaner” in the sense that you have a more orderly universe of correlation coefficients, sensible eigenvalues and a variable plot that does not get swamped by the size effect of the variables. To use it, however, you have to be certain that you do not include variables that hold nonessential information, variables that just provide random noise or variables that are far from normality. Through the standardisation process all variables become of equal importance. The best you can do is to use both of the methods, experiment with them and compare the results. In this way you will also learn something about your variables.

Example using measurement data on 66 Early Neolithic pots

The data for this example is taken from Koch 1998. In this study a total of 153 complete Funnel beakers were measured and drawn, and the profiles were subsequently scaled to the same height and visually compared (Koch 1998 p. 67 ff). Based on this comparison nine different shape modes were separated covering the whole of the Early Neolithic and the first half of the Middle Neolithic. Many of the shape modes are very close to each other, but characteristic types of decorations help to separate them. A PCA on all pots shows a confusing, mixed spread of pots from various shape groups. It is however possible to see that there is a pattern among the early shape groups and the later shape groups isolated. This was why PCA’s for the early and the late material originally were run separately (Koch 1998 p.71 ff.), and why the material selected here for the example only comprises 66 pot of shape groups 0, 1, 2 and 3.





The original measurements were taken as coordinates to characteristic points of the pot profile following the scheme shown above to the left. For this example these coordinates have been recalculated into a number of characteristic measurements as shown above to the right. Apart from making it easier to interpret the variable patterns in the analysis, this kind of measurement scheme also makes it possible to include fragmented material, say the neck part of pots, and then only analyse the neck variables.

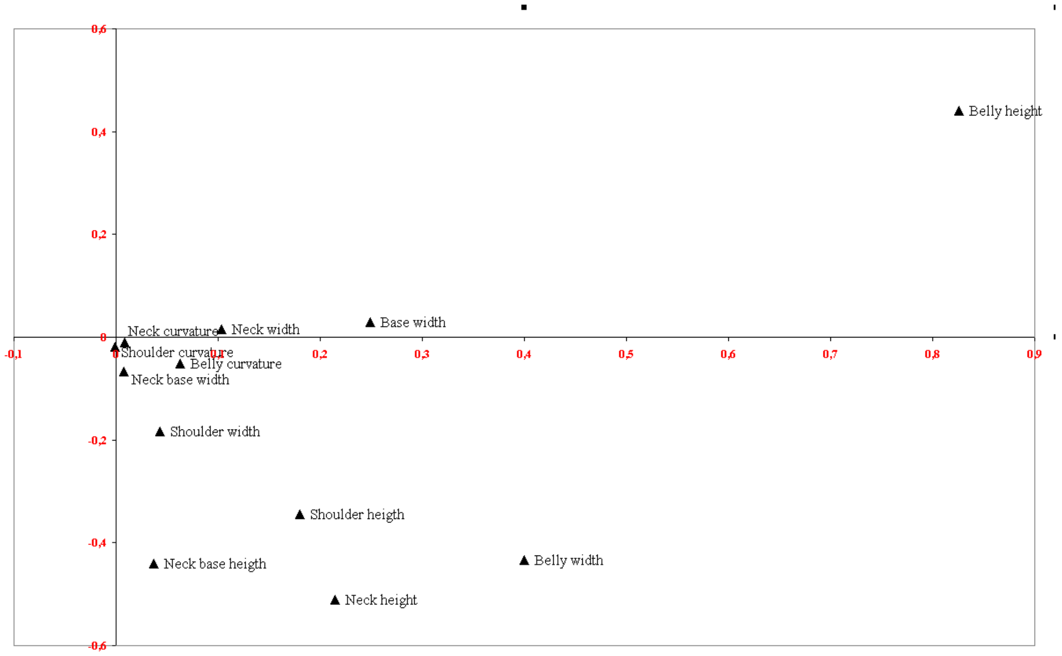
The characteristic measurements calculated are all in mm and should be directly comparable. There are, however, considerable differences between the measuring span of the variables. Especially the curvature measurements and neck base width are numerically small and with a limited span. It is clearly to be expected that a PCA based on the correlation matrix will be distorted by the variation in these variables unless they co-vary very systematically with some of the larger variables. It is therefore the obvious choice to start out with a PCA based on the covariance matrix.

Covariance matrix

	Base width	Belly height	Belly width	Belly curvature	Shoulder heigth	Shoulder width	Shoulder curvature	Neck base heigth	Neck base width	Neck height	Neck width	Neck curvature
Base width	2,35											
Belly height	5,49	18,73										
Belly width	2,25	8,41	5,41									
Belly curvature	0,31	1,34	0,84	0,15								
Shoulder heigth	1,44	3,55	2,14	0,34	1,89							
Shoulder width	0,36	0,75	0,62	0,09	0,57	0,25						
Shoulder curvature	0,00	-0,02	0,02	0,00	0,02	0,02	0,00					
Neck base heigth	0,34	0,38	1,08	0,12	0,31	0,21	0,03	1,22				
Neck base width	0,07	0,11	0,17	0,02	0,07	0,04	0,00	0,18	0,04			
Neck height	1,57	4,31	2,94	0,43	1,69	0,56	0,03	0,63	0,13	2,58		
Neck width	0,61	2,28	1,15	0,19	0,48	0,08	0,00	-0,07	0,00	0,73	0,48	
Neck curvature	0,04	0,17	0,11	0,02	0,04	0,01	0,00	0,02	0,00	0,08	0,03	0,01

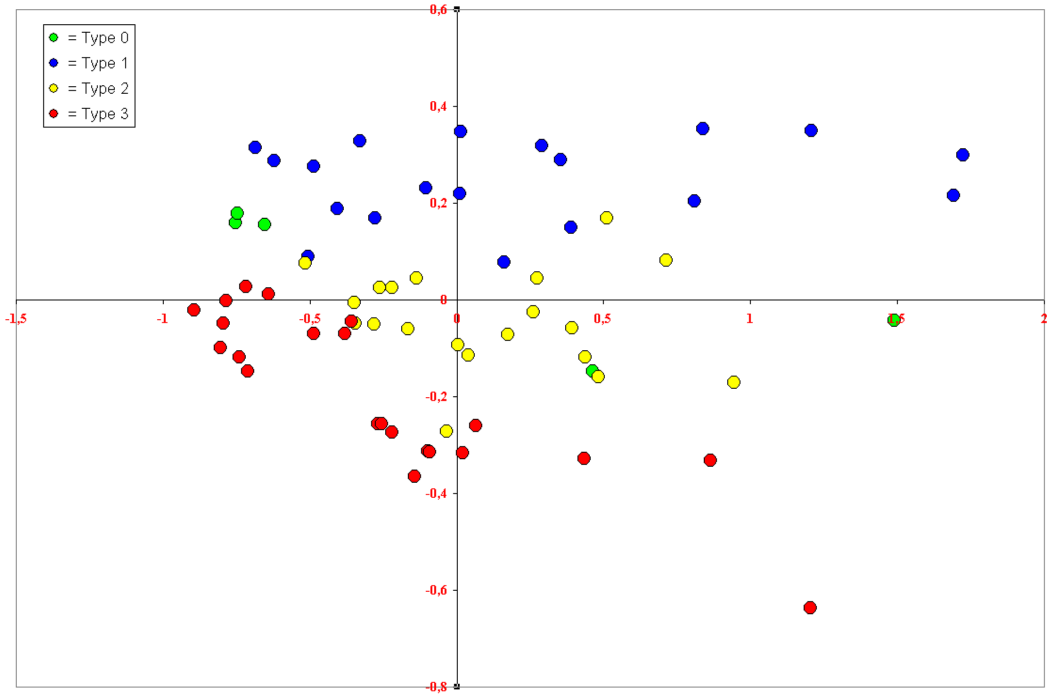
Covariance matrix between 12 measurement variables based on data from 66 pots. The variables are not weighted.

Looking at the covariance matrix we find, as in the previous example, that the size of the coefficients in the diagonal cells quite clearly reflects the span of the variables. It is not as marked as with lance heads, but looking at the rest of the coefficients, seeing that very few are negative and larger values primarily occur where variables with bigger spans combine, we can clearly expect the variable plot to sort the variables according to size of value span.



Biplot of variable loadings from a PCA based on covariance coefficients. Data consist of 66 pots measured by 12 variables. The variables are not weighted.

This is also what happens. Indeed, if you compare the size of the coefficients on the diagonal with the placement of the variables along the first principal component of the variable plot, you will find that *Belly height*, which has the highest value, lies to the right, and *Shoulder curvature*, which has the smallest value, lies to the left. The rest are spread in between in rank order of size with a distance between them that reflects their approximate size difference.

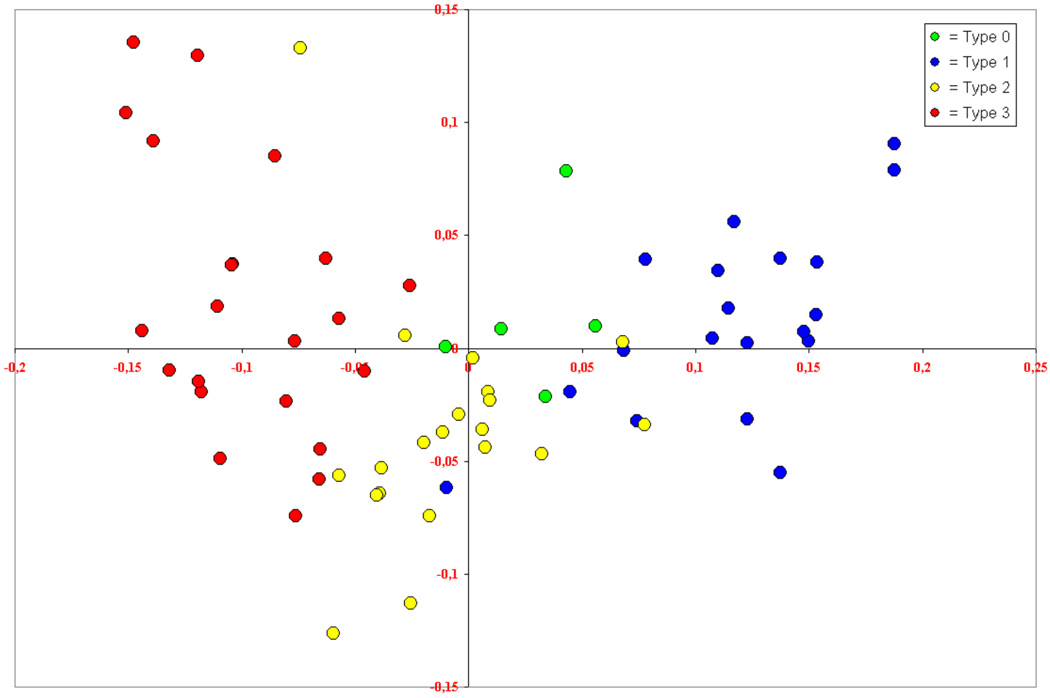


Objects plot of a PCA based on covariance coefficients. Data consist of 66 pots measured by 12 variables. The variables are not weighted.

When we look at the object plot we find that types 1, 2 and 3 are separated, but only in the second principal component and not the first, where you would expect the major separation to occur. In stead pots from all three types are strewn out along the component. The reason becomes immediately obvious, when you check the information on the individual pots. Those furthest to the right are the biggest pots and those furthest to the left are the smallest ones. The first principal component simply sorts the pots according to size.

This size sorting happens very often in a PCA if the magnitude of the variable values depends on the size of the objects. It did not happen with the lance heads because size itself in this case seems to be a constituting element in the types, but with ceramic pots it is an entirely different matter. The shape of a pot is more or les independent of its size. Pots with the same shape can easily be made in very differing sizes depending on functional needs. The problem with the size element in a PCA arises because the measurements taken to outline the shape of the pots at the same time are measurements of size.

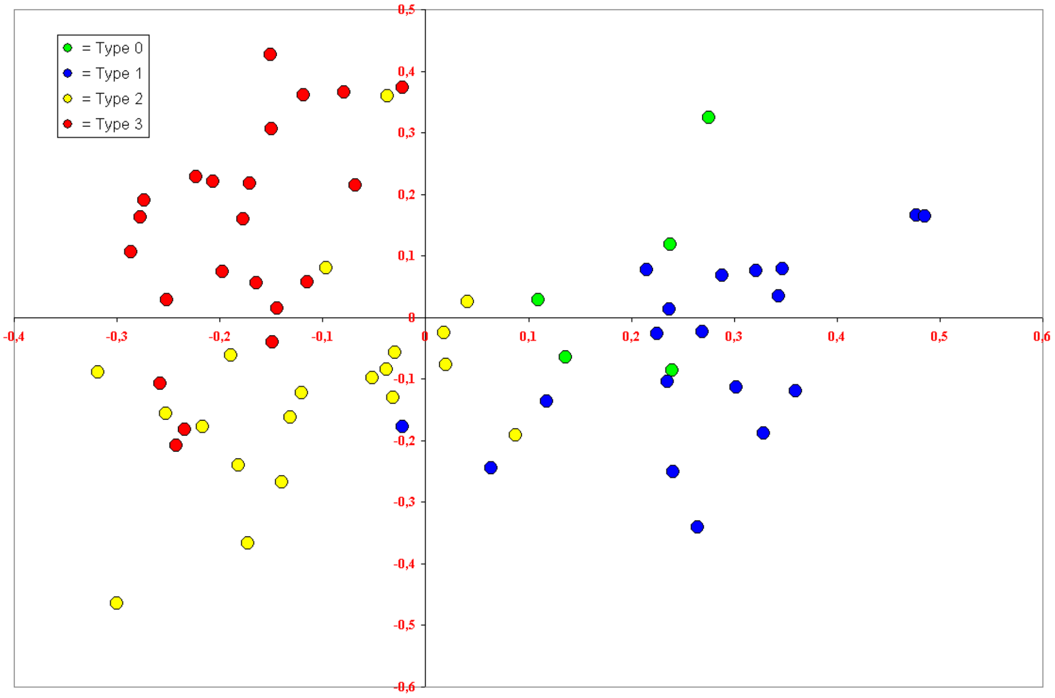
To avoid size sorting in the first principal component you have either to devise some size independent measures - indexes and angles (creating a variety of other problems), or much better create a weighting factor for the objects that removes the size factor. One obvious solution with pots would be to use the volume as a weighting factor (1 divided with the cube root of volume would be an appropriate weighting factor). Much simpler and just as efficient is a factor for each object based on the sum of all measurements for the object. This would be useable for analyses of both complete pots and analyses of parts of pots, where the volume information for the part analysed may not be available. The weighting factor used in the following is (10/sum of measurements) for each pot. In CAPCA weights have to lie between 1 and 0. The factor 10 is here chosen because it nicely balances the weights within this interval. Anyway, you will have to choose a factor that is equal to or smaller than the smallest sum. Otherwise you will get weights larger than 1.



Objects plot of a PCA based on covariance coefficients. Data consist of 66 pots measured by 12 variables. The variables are weighted to eliminate size as a discriminating factor.

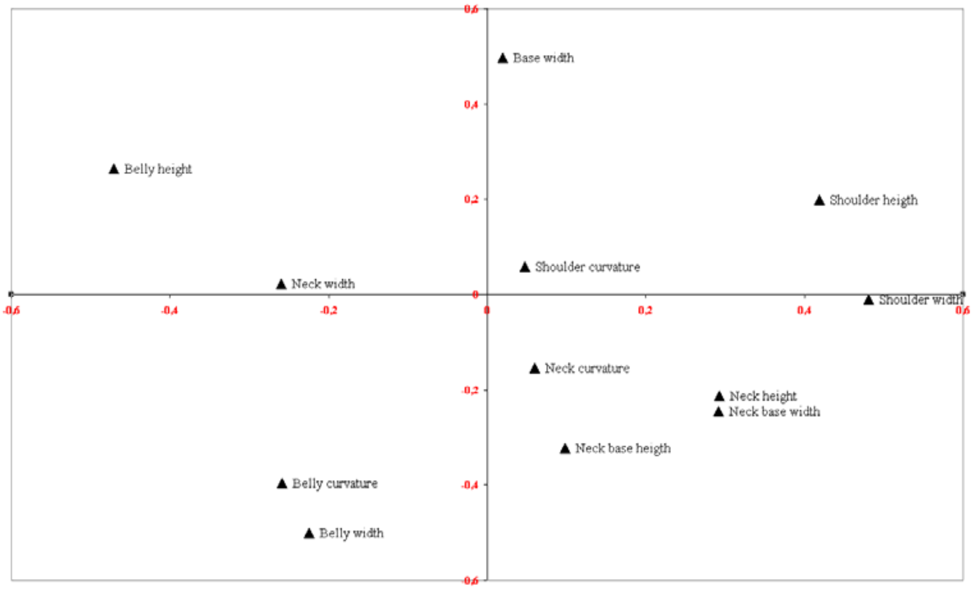
Clearly the use of weights to eliminate the effects of size on objects work. The first principal component now shows variations in shape separating Type 1, 2 and 3. There are no distinct groupings, however. Rather a continuous development is suggested, which would be in agreement with a claimed typological development from Type 1 to Type 3 held by some scholars. The position of Type 0, claimed to be the oldest type, begs for an archaeological explanation, however.

The next step is to run a PCA based on the correlation matrix using the same weights.



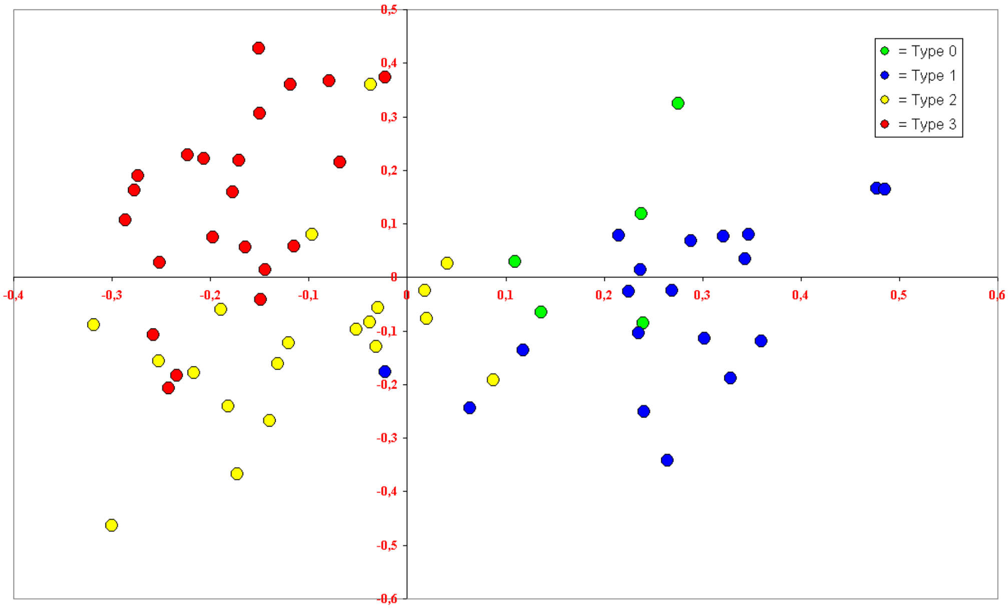
Objects plot of a PCA based on correlation coefficients. Data consist of 66 pots measured by 12 variables. The variables are weighted to eliminate size as a discriminating factor.

The result shows the same general trend as that of the covariance matrix with type 1 at one end type 2 in the centre and type 3 at the other end of the first principal component. However, there are also differences. Type 0 now seems more adjacent to Type 1 and there is a tendency for a break in the middle between Type 1 and type 2 and 3. This would tend to be in line with the suggestions by other scholars, who claim that there is a major difference between type 1 and its cultural milieu on the one hand and type 2 and 3 and their cultural milieu on the other.



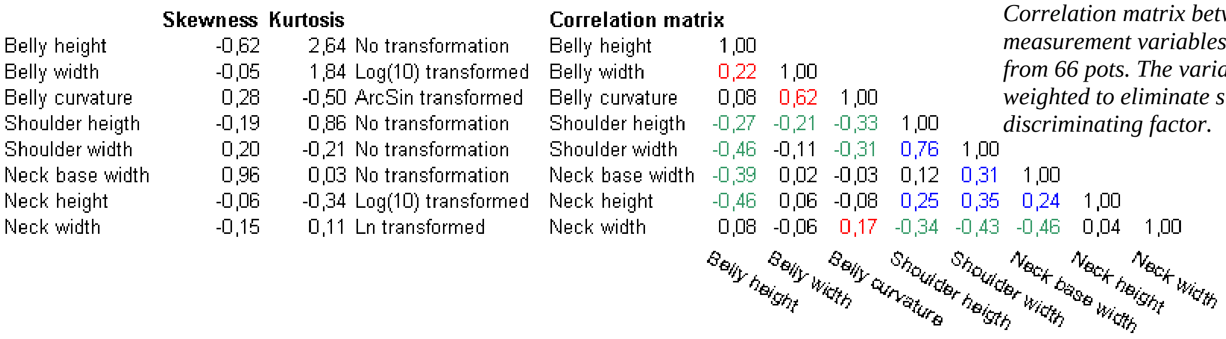
Biplot of variable loadings from a PCA based on correlation coefficients. Data consist of 66 pots measured by 12 variables. The variables are weighted to eliminate size as a discriminating factor.

If we look at the variable plot we can see that there is an opposition on the first principal component between *Belly height*, *Neck width*, *Belly curvature* and *Belly width* to the left and *Shoulder width*, *Shoulder height*, *Neck height* and *Neck base width* to the right. It is clearly these variables that condition the potential bipartition of the pots. In the middle are four variables that do little to this division except possibly confuse it. Potentially it could give a clearer picture if we left out *Base width*, *Shoulder curvature*, *Neck curvature* and *Neck base height*.



Objects plot of a PCA based on correlation coefficients. Data consist of 66 pots measured by 8 variables. The variables are weighted to eliminate size as a discriminating factor.

The break between the two groupings has become slightly more accentuated, but overall the picture has not changed, and the four excluded variables had thus no significant influence on the result. If we turn to the statistics and the variable plot we may try to interpret the result a little closer.

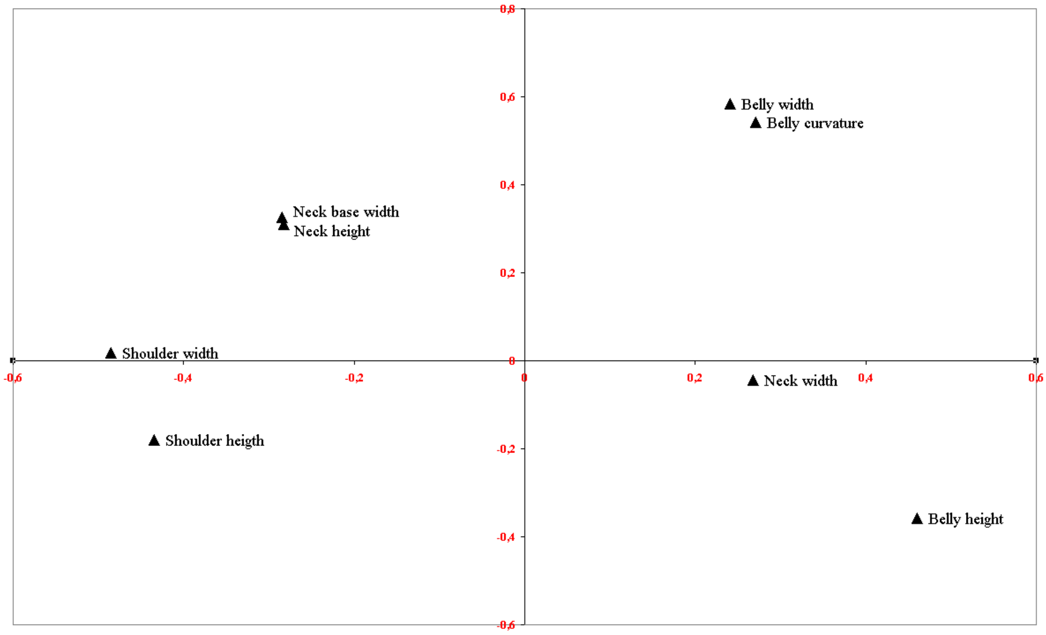


	Skewness	Kurtosis	
Belly height	-0,62	2,64	No transformation
Belly width	-0,05	1,84	Log(10) transformed
Belly curvature	0,28	-0,50	ArcSin transformed
Shoulder heigth	-0,19	0,86	No transformation
Shoulder width	0,20	-0,21	No transformation
Neck base width	0,96	0,03	No transformation
Neck height	-0,06	-0,34	Log(10) transformed
Neck width	-0,15	0,11	Ln transformed

Correlation matrix	Belly height	Belly width	Belly curvature	Shoulder heigth	Shoulder width	Neck base width	Neck height	Neck width
Belly height	1,00							
Belly width	0,22	1,00						
Belly curvature	0,08	0,62	1,00					
Shoulder heigth	-0,27	-0,21	-0,33	1,00				
Shoulder width	-0,46	-0,11	-0,31	0,76	1,00			
Neck base width	-0,39	0,02	-0,03	0,12	0,31	1,00		
Neck height	-0,46	0,06	-0,08	0,25	0,35	0,24	1,00	
Neck width	0,08	-0,06	0,17	-0,34	-0,43	-0,46	0,04	1,00

Correlation matrix between 8 measurement variables with data from 66 pots. The variables are weighted to eliminate size as a discriminating factor.

The variables are generally well behaved with respect to Skewness and Kurtosis. The correlation coefficients are not particular high, and their patterning can be a little difficult to see at first, but there are two groups of variables with mutual positive correlation. The one consist of *Belly height*, *Belly width*, *Belly curvature* and partly *Neck width* (red), the other of *Shoulder height*, *Shoulder width*, *Neck base width* and *Neck height* (blue). The coefficients between members from these two groups show marked negative correlation (green). Clearly this pattern lies behind the splitting of the pots in two groups. How it does can best be seen from the variables plot.



Biplot of variable loadings from a PCA based on correlation coefficients. Data consist of 66 pots measured by 8 variables. The variables are weighted to eliminate size as a discriminating factor.

In the variables plot the first group (red) lies to the right on the first principal component, which is the side where Type 0 and Type 1 pots are placed in the objects plot. The other group (blue) lies to the left on the first principal component, which is the side where Type 2 and Type 3 pots are placed in objects plot. The negative correlation between members of the two groups is reflected in their position on both sides of zero. An interpretation in general terms would be that Type 0 and 1 pots have a high and wide, curved belly, a small insignificant shoulder and a low, flaring (wide) neck. Type 2 and 3 pots on the other hand have a low and not very wide belly, a pronounced shoulder and a generally high, non-flaring neck with a clear tendency for a narrowing of the lower part. If we focus on the second principal component, we find from the objects plot that Type 2 pots lies on the negative side of the component and Type 3 pots lie on the positive side. Judging from the variables plot we can suggest that Type 3 pots have generally higher necks than Type 2, whereas Type 2 pots have generally more pronounced shoulders than Type 3. The high position on the second principal component of Belly width and Belly curvature probably reflects that the small group of Type 3 pots and one Type 2 pot that separates themselves in the top centre of the objects plot has fairly wide and curved bellies.

In this example the PCA on a correlation matrix came out just as good as or better than the one on the covariance matrix. The reason must be that in this case the variables with a very small value span are just as meaningful and informative as those with a large value span. They do not become noise emitters when normalised. There is thus no way in which we can decide in advance if we should choose a covariance or a correlation matrix. We have to argue from the nature of the variables combined with results of actual analyses, which one we should prefer.

CORRESPONDENCE ANALYSIS

Correspondence Analysis takes counts rather than measures as input. Counts are by definition positive integers with zero being the state of no occurrence. When counts are presented in a table we call this a contingency table. The first thing to note about a contingency table is that by definition all entries are on the same scale (counts are counts), which means that in contrast to measurement data we can perform calculations across both variables and objects, and not just across the variables. Thus creating sums of counts on objects across variables (row sums) is just as meaningful as creating sums of counts on variables across objects (column sums).

To work with a contingency table we need to have a notion of what constitute structure in the table, and by the same token how lack of structure should be defined. The row and column sums in a way defines the basic content of the table, and we can claim that the content of a table is unstructured if the individual cell values is a mere random reflection of the row and column sum values. This so called randomized table or table of expected values can simply be obtained for each cell by multiplying the corresponding row sum and column sum and divide the result with the total number of counts in the table.

The structure of contingency tables can then be seen as deviations between observed and expected cell values. The actual measure used is calculated as ((observed cell value - expected cell value) / square root of expected cell value). This is in fact equal to the terms used in chi-square tests.

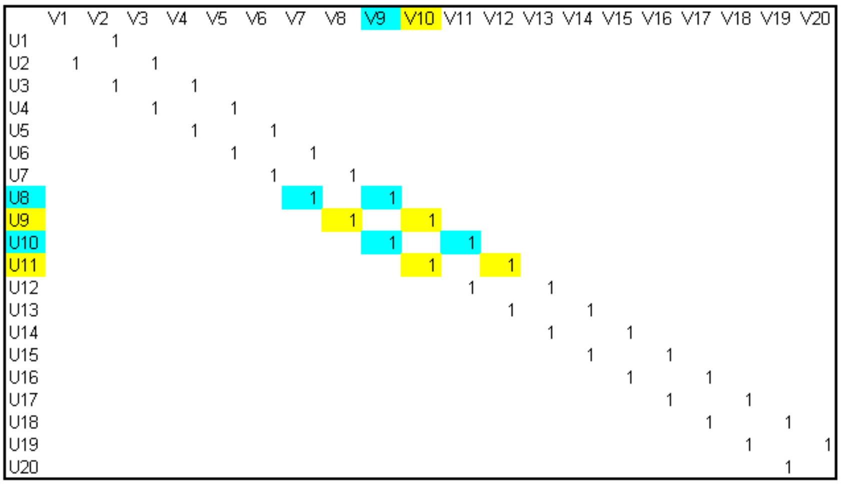
When performing an orthogonal regression on this table it will be the patterns of deviation above and below the expected values that will form the covariance patterns leading to the separation of a set of new axes. This new set of axes created by orthogonal regression is referred to as *principal axes* (rather than principal components to avoid confusion with PCA). The principal axes are ranked in such a way that the first axis has the largest representation of the total variation in the data set, the second axis the second largest part, etc. They may be viewed as a new set of variables substituting the original variables, and in doing so they retain the total amount of variability in the data, but represent it in a different more structured way.

As in PCA we can speak of loadings and scores, but it is problematic to do so. First of all it is not constant in CA what should technically be considered variables and objects. This is an important difference to PCA where variables are treated differently than objects. In CA there is no difference and the smallest dimension of the dataset is computationally considered to be the variables. Secondly because of the equality, or symmetry if you wish, of variables and objects, loadings (and scores) are not scaled in the same way as in PCA, and there is thus no direct equivalence with correlation coefficients. Consequently, it is confusing to speak of loadings and scores. Because they are scaled equally, to allow them to be plotted together in one diagram, it is customary to speak of *variable coordinates* and *object coordinates* in stead.

As in PCA an *eigen value* is associated with each principal axes, but the explanation of the value is not as straight forward as with PCA. The total sum of eigen values, which in CA is called the *inertia* of the data set, is not equal to the number of variables, but usually much smaller due to scaling. The part of the total inertia represented by the eigen values of each principal axis does, however, show the part of the total variation covered by the individual axes exactly as in PCA.

There are no immediate concerns before running a CA. Just press the button to run the analysis. Understanding the result, however, is not straight forward, and in many aspects it can become more complicated than with a PCA. Indeed you may have to run a lot more analyses, where you step by step alter the input. The alterations to the input may consist of weighting of objects and/or variables and of omission of either objects or variables. To decide what to do, calls for an understanding of the output in both graphical and numerical format.

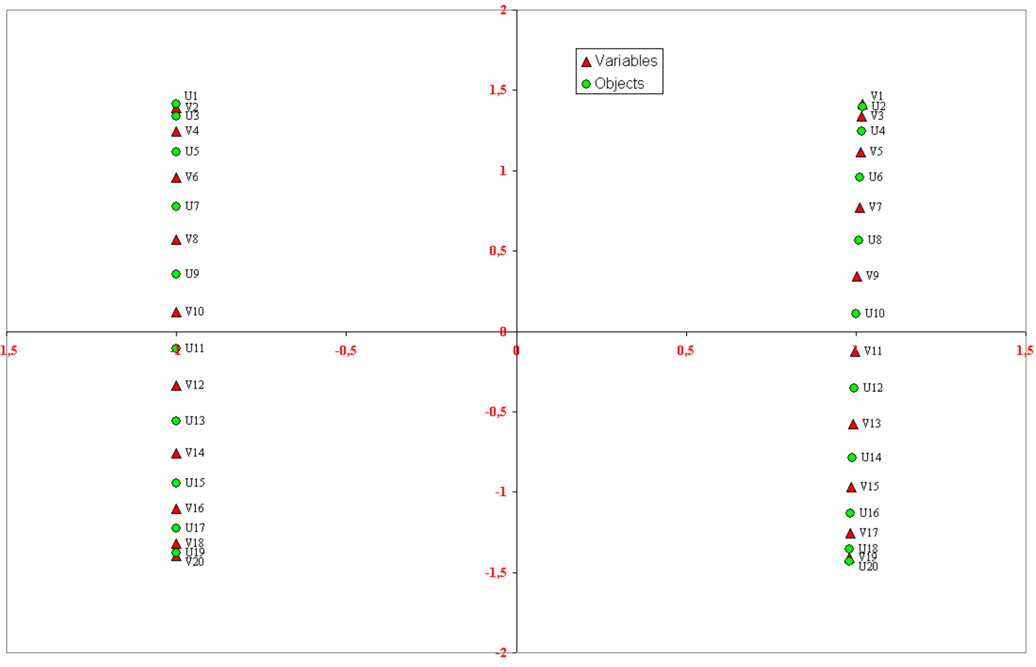
The first thing to emphasize, as already stated, is that variables and objects can be presented together in the coordinate system formed by the principal axes. Further in doing so the position of the variables in the plot is directly interpretable in relation to the objects and vice versa. We can visually inspect and interpret variables and objects in one single plot. To gain a better understanding of visual interpretation we will look at a few artificial examples with idealised matrixes. We start out with the matrix below, where the cells on each side of the diagonal filled is with 1’s, and the rest is filled with 0’s (blanks are always 0 in a CA).



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
U1		1																		
U2	1		1																	
U3		1		1																
U4			1		1															
U5				1		1														
U6					1		1													
U7						1		1												
U8							1		1											
U9								1		1										
U10									1		1									
U11										1		1								
U12											1		1							
U13												1		1						
U14													1		1					
U15														1		1				
U16															1		1			
U17																1		1		
U18																	1		1	
U19																		1		1
U20																			1	

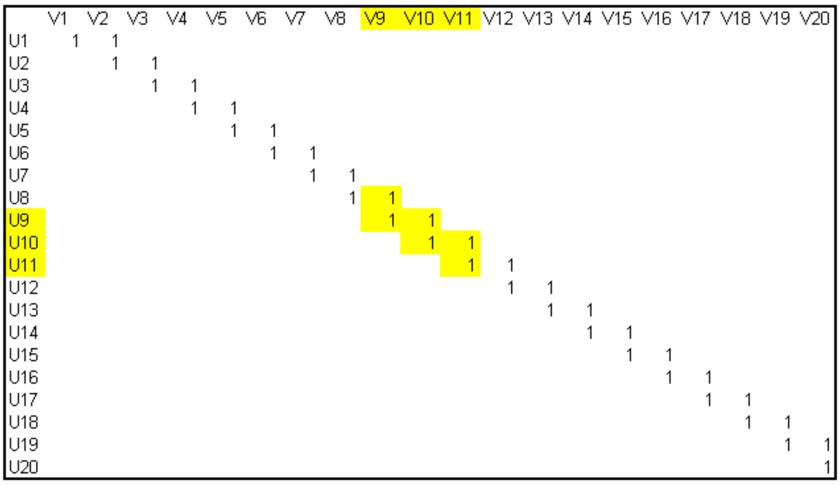
Idealised 20 by 20 matrix with two independent sets of objects and variables. Within each set the objects and variables are linked together in a chain of shifting objects and variables.

This matrix is quite interesting as the two rows of 1’s form a peculiar pattern of what goes with what. Object U8 is linked with object U10 through variable V9, and object U9 is linked with objects U11 through variable V10. Objects U8 and U10, however, are not linked with U9 and U11 in any way. Nor are variable V9 and Variable V10 linked to each other in any way. This pattern continues throughout the matrix dividing objects and variables in two sets with the same number of objects and variables in each set. There is no connection between the two sets, but within each set there is a systematic relationship between objects and variables as each object is linked to the next object through one variable and each variable is linked to the next variable through one object. The graphical representation of this is shown in the following plot of the first two principal axes.



CA of an idealised 20 by 20 matrix with two independent sets of objects and variables. Within each set the objects and variables are linked together in a chain of shifting objects and variables. Combined plot of 1. and 2. principal axis.

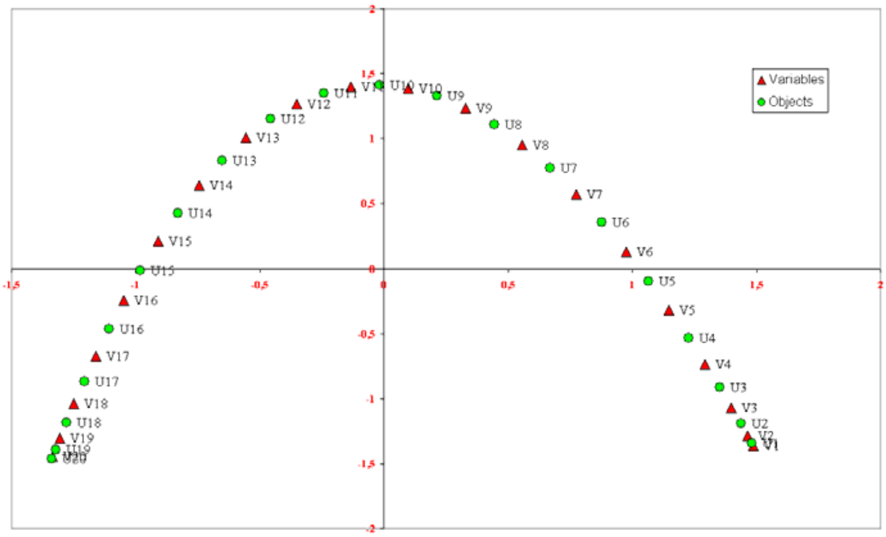
We find that the two sets are placed at each end of the first principal axes reflecting the lack of connection between them. Within each set there is a very systematic layout that reflects that each object is linked by exactly one variable to the next object, and that each variable is linked by exactly one object to the next variable. All together they form a chain of shifting objects and variables that is laid out as a straight line on the second Principal Axis with a constant distance between objects and variables (except for the edge effects of the matrix). You can interpret this layout directly in simple terms of closeness between objects and variables as reflected by their combinations.



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
U1	1	1																		
U2		1	1																	
U3			1	1																
U4				1	1															
U5					1	1														
U6						1	1													
U7							1	1												
U8								1	1											
U9									1	1										
U10										1	1									
U11											1	1								
U12												1	1							
U13													1	1						
U14														1	1					
U15															1	1				
U16																1	1			
U17																	1	1		
U18																		1	1	
U19																			1	1
U20																				1

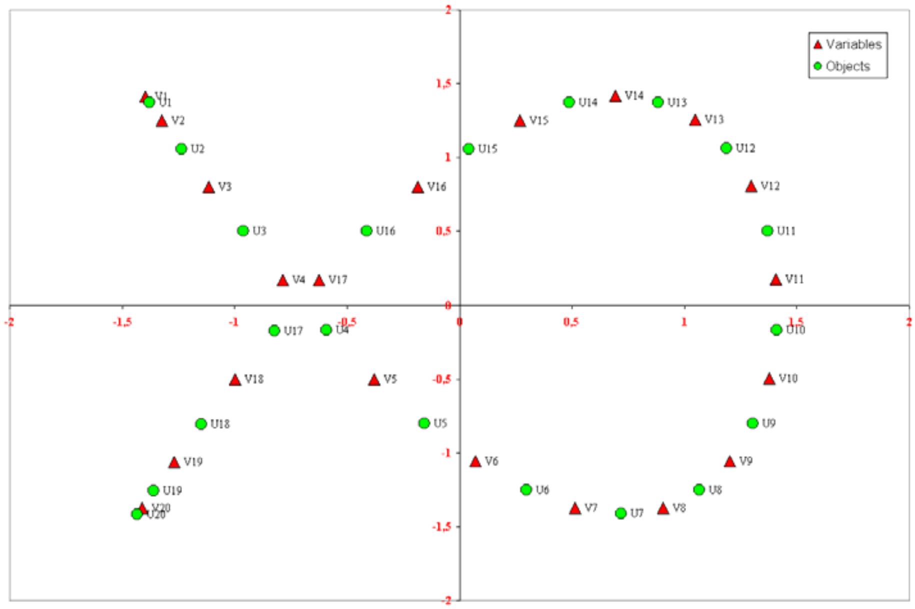
Idealised 20 by 20 matrix with objects and variables that are linked together in one chain of shifting objects and variables.

Now let us try to see what happens when we fill the diagonal and the cells adjacent to the diagonal to one side with 1's, while the rest of the matrix is filled with 0's. We see that V9 is linked to U9, which is linked V10, which is linked to U10, which is linked to V11, which is linked U11, etc. Thus all objects and variables form a single chain with shifting objects and variables. This results in the following plot.



CA of an idealised 20 by 20 matrix with objects and variables that are linked together in one chain of shifting objects and variables. Combined plot of 1. and 2. principal axis.

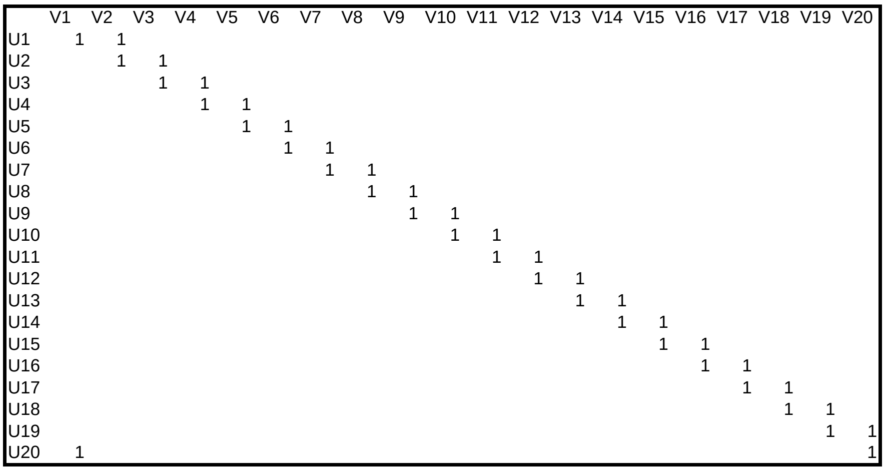
The arc shaped layout is characteristic for data with a pattern of continuity between objects and variables. That is when you move across the objects there is a gradual and systematic replacement of variables and vice versa. This is also the criteria for seriation, and the perfect arc shaped layout indicates that the data meets the criteria for at perfect seriation. Again we see the equally spaced objects and variables as would be expected from their systematically chained relationship. Why it shows up as an arc may not seem evident. It has to do with multidimensionality, though. The plot you see is two dimensional, but you must not think of it as two dimensional. It is a line that passes through multidimensional space, and the arced layout is a result of a projection into two dimensions. It is much like looking at the maps of international flight destinations that you find in the flight magazines of any plane. They all form arcs, not because the planes fly a detour, but because the shortest route around the globe appears as a curved line on the two-dimensional projection of a map. The objective here, however, is not the shortest line, but a line along which the objects and variables are evenly and maximally spread. On the first and second axis this is not the case. The distribution is denser in the middle and towards the end. If you include the third axis (below) you can see why. Over three axes we are not speaking of an arc, but a spiral.



CA of an idealised 20 by 20 matrix with objects and variables that are linked together in one chain of shifting objects and variables. Combined plot of 2. and 3. principal axis.

You can find a very detailed discussion of the behaviour of different types of continuity patterns in data sets that can be seriated in Jensen & Nielsen 1997.

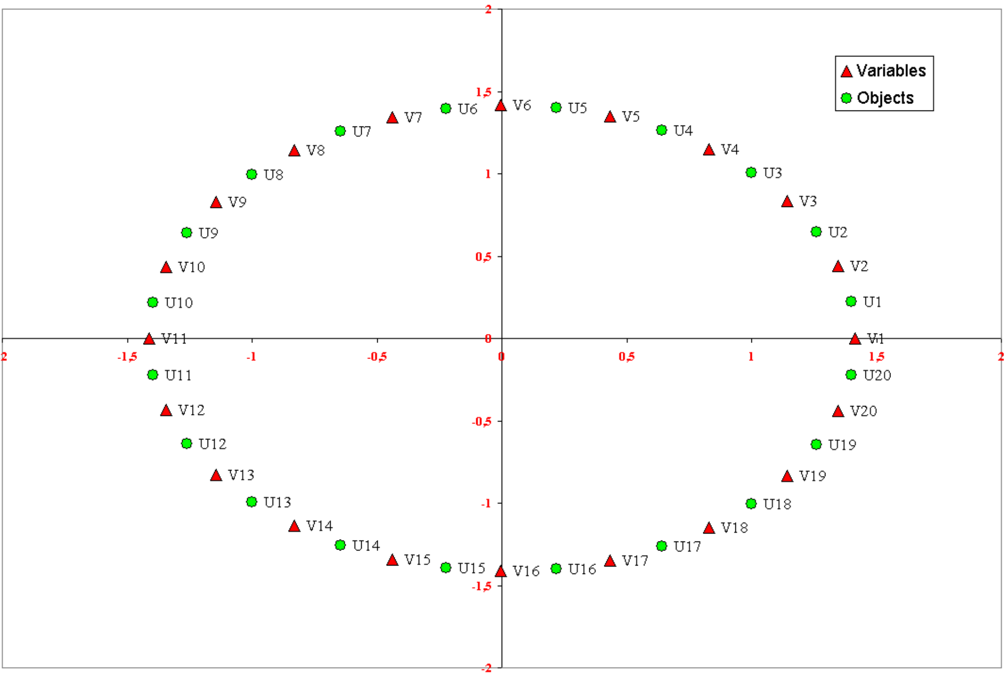
V1 and U20 marks the ends of the chain being linked in one direction only. What will happen if we link the two together by inserting 1 in the cell that combines V1 and U20, and hence create a continuous circular chain?



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
U1	1	1																		
U2		1	1																	
U3			1	1																
U4				1	1															
U5					1	1														
U6						1	1													
U7							1	1												
U8								1	1											
U9									1	1										
U10										1	1									
U11											1	1								
U12												1	1							
U13													1	1						
U14														1	1					
U15															1	1				
U16																1	1			
U17																	1	1		
U18																		1	1	
U19																			1	1
U20	1																			1

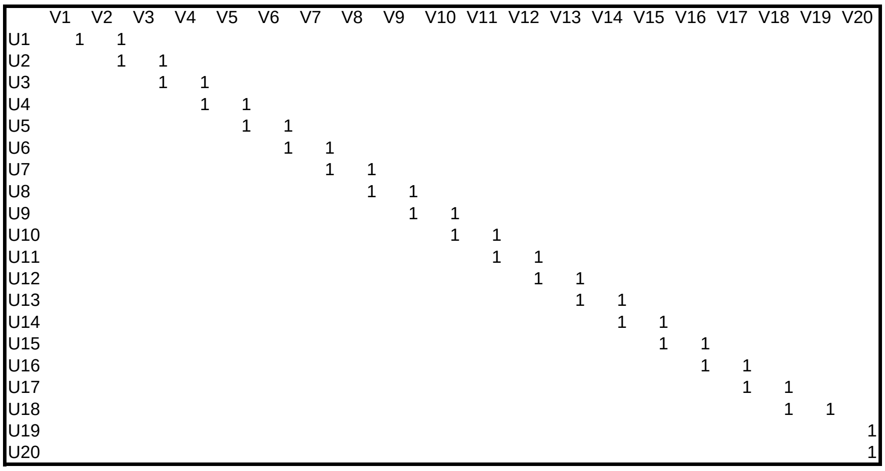
Idealised 20 by 20 matrix with objects and variables that are linked together in one chain of shifting objects and variables, and where the ends of the chain have been "fused" together.

Logically the result should be a circle, and it is.



CA of idealised 20 by 20 matrix with objects and variables that are linked together in one chain of shifting objects and variables, and where the ends of the chain have been "fused" together. Combined plot of 1. and 2. principal axis.

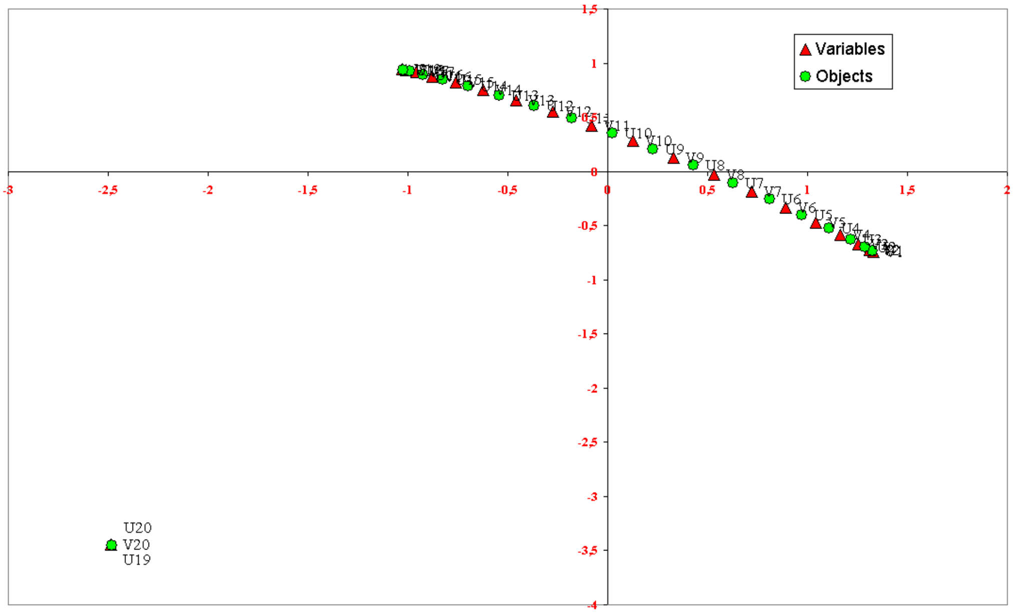
Now what happens if we break the chain by setting 0 between U19 and V19, and thus isolating U19, U20 and V20?



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
U1	1	1																		
U2		1	1																	
U3			1	1																
U4				1	1															
U5					1	1														
U6						1	1													
U7							1	1												
U8								1	1											
U9									1	1										
U10										1	1									
U11											1	1								
U12												1	1							
U13													1	1						
U14														1	1					
U15															1	1				
U16																1	1			
U17																	1	1		
U18																		1	1	
U19																				1
U20																				1

Idealised 20 by 20 matrix with objects and variables that are linked together in one chain of shifting objects and variables, apart from two objects and one variable that has been isolated through a break in the chain.

We find that U19, U20 and V20 are placed together in one corner, while the remainders, still forming a chain, are placed in the opposite corner as a slightly curved line. U19, U20 and V19 are what are referred to as outliers – objects or variables that either, as in this example, is more or less uncorrelated with the rest of the material, or displays excessive values that set them apart from the rest of the material. The latter situation occurs only in connection with counts of numerous occurrences, and will be dealt with in the examples below.



CA of idealised 20 by 20 matrix with objects and variables that are linked together in one chain of shifting objects and variables, apart from two objects and one variable that has been isolated through a break in the chain. Combined plot of 1. and 2. principal axis.

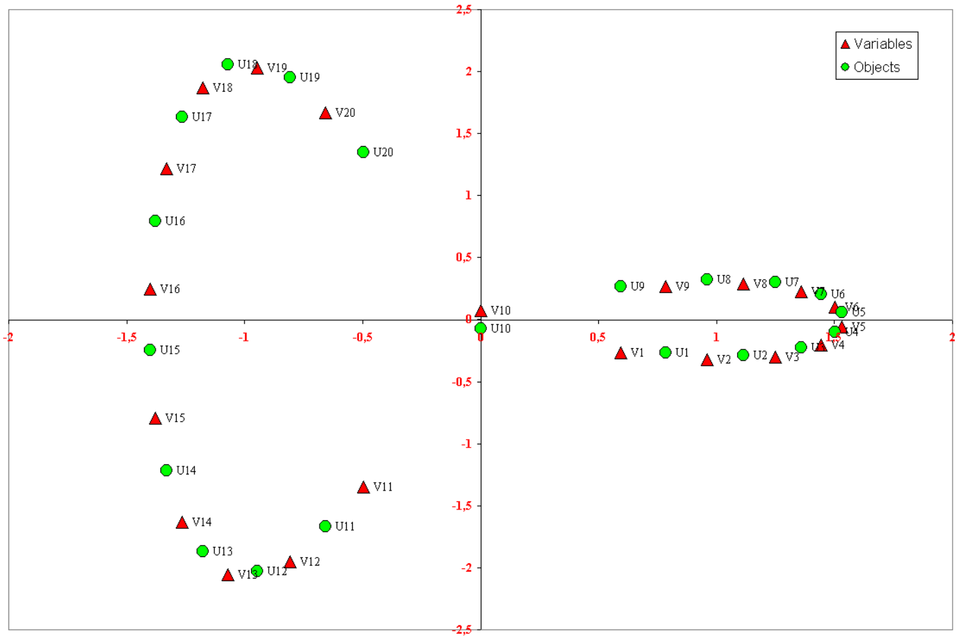
To handle outliers there are only two possibilities: either you remove them from the analysis or you use weights to change their behaviour. With the outlier example above, where a complete break in continuity of the material exists, removal is the only option. You simply note that U19, U20 and V19 are not related to the rest of the material in any way, and then remove them. In other situations you can either remove or use weights. Weighting is an essential part of CA's, but you have to carefully consider when and how to use it. In connection with the examples below weighting will be discussed in more detail.

Finally, what will happen if we introduce objects and variables with a constant appearance throughout the data? In the following variable V10 appears in all objects, and object U10 scores on all variables.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
U1	1	1								1										
U2		1	1							1										
U3			1	1						1										
U4				1	1					1										
U5					1	1				1										
U6						1	1			1										
U7							1	1		1										
U8								1	1	1										
U9									1	1										
U10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
U11										1	1	1								
U12										1		1	1							
U13										1			1	1						
U14										1				1	1					
U15										1					1	1				
U16										1						1	1			
U17										1							1	1		
U18										1								1	1	
U19										1									1	1
U20										1										1

Idealised 20 by 20 matrix with objects and variables that are linked together in one chain of shifting objects and variables, apart from one object that appear with all variables and one variable that appear in all objects, and thus acts as constants.

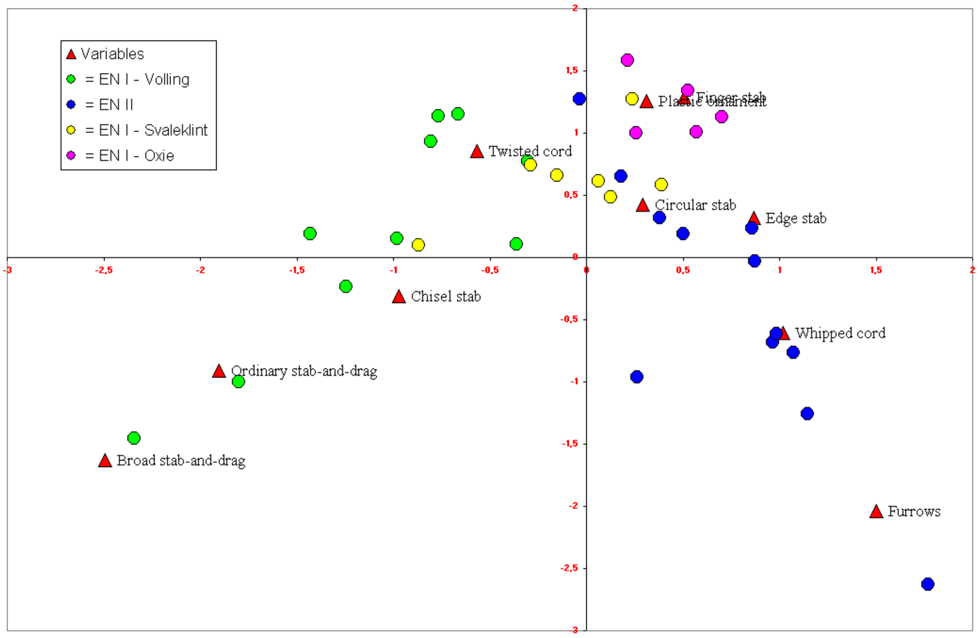
The result is that V10 and U10 are placed together in the centre of the plot with the remaining objects and variables stretched out in arcs on both sides of the centre. The seriation totally collapses, and had it not been for the very systematic distribution along the diagonal, the objects and variables would merely have formed a loosely clustered group around the centre. You should be very wary of variables with a constant appearance in your data. Constant variables can ruin the result of any CA, whether you are looking for a seriation, clusters, or any other form of structure. In CAPCA you can use the information on the sheet *Matrix output* to track them down. A constant variable is one that has a high frequency of non zero cells in connection with presence absence data (look at the table *Data as analysed*) and one that has a uniform value profile across many objects in connection with counts (look at the table *Data sorted by rank on first principal axis*).



CA of idealised 20 by 20 matrix with objects and variables that are linked together in one chain of shifting objects and variables, apart from one object that appear with all variables and one variable that appear in all objects, and thus acts as constants. Combined plot of 1. and 2. principal axis.

Example using counts of technical elements on rim shards from pottery
The material used in this example is adopted from Madsen & Petersen 1984. It is counts of 10 different technical elements in the decoration on rims of Early Neolithic pottery from 34 settlement sites. The size of the sites, or rather the extent of the excavations, varies considerably, and the number of decorated rim shards for each site varies accordingly. Thus the largest site has 304 decorated shards, the smallest six.

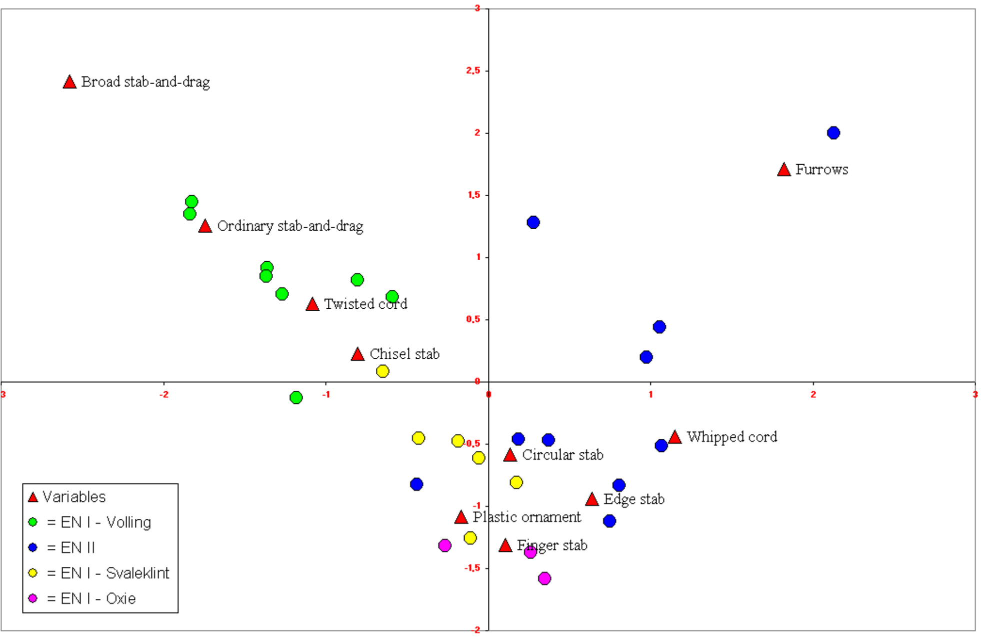
The sites can be classified according to their regional and chronological groupings using traditional typo-chronological criteria. This has led to a division into three partly regional EN I groups and one EN II group.



CA of the occurrence of ten different technical rim decoration elements in 34 settlements. Objects and variables are not weighted. Combined plot of 1. and 2. principal axis.

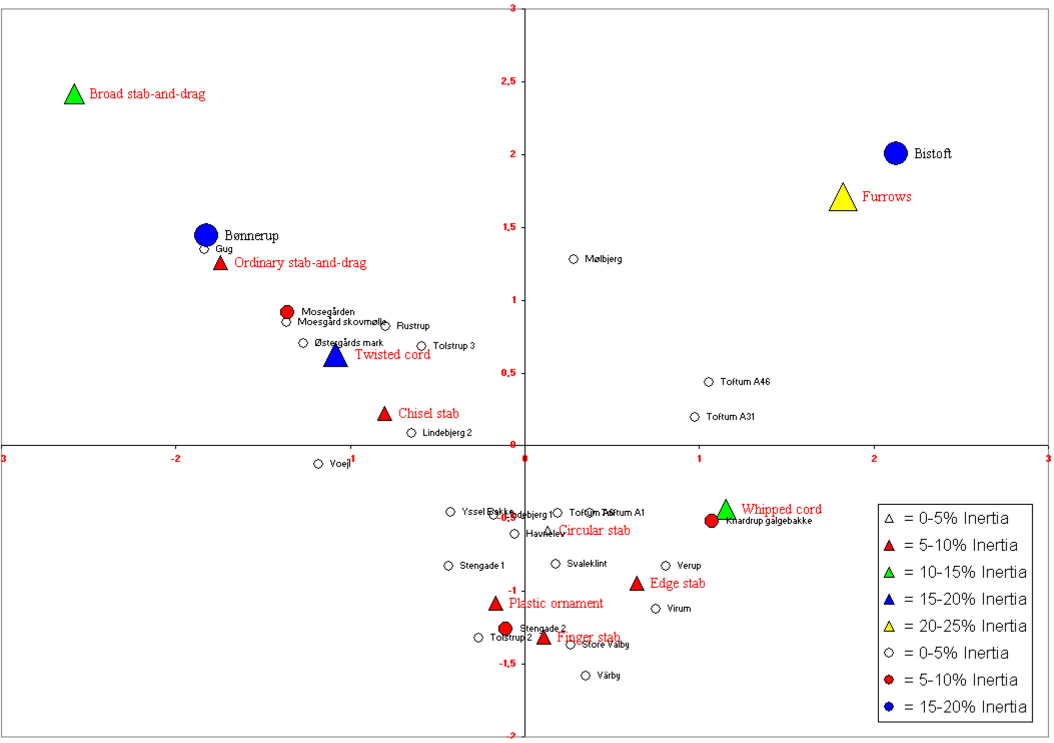
Looking at a first plot of objects and variables together, we can see that the four different groups have been separated, at least partially. It can immediately be seen that there is a tendency for outliers in the lower left and right hand corners. However, before attending to these we should set a standard for the objects. There are seven sites that have less than ten counts of elements. We exclude these from the analysis as too uncertain. Then there are six sites with more than 100 counts of elements. To avoid that these by sheer number becomes too influential we weight them down to a sum of 100, in reality changing their counts to percentages.

With the renewed analysis we find (below) that the change does not make much difference to the layout (except that the values on the second principal axis have been mirrored), but we are now certain that the sites are as comparable as we can make them. We could of course set the minimum sum higher, but that would quickly cut down the number of sites in the analysis and make it of little use.



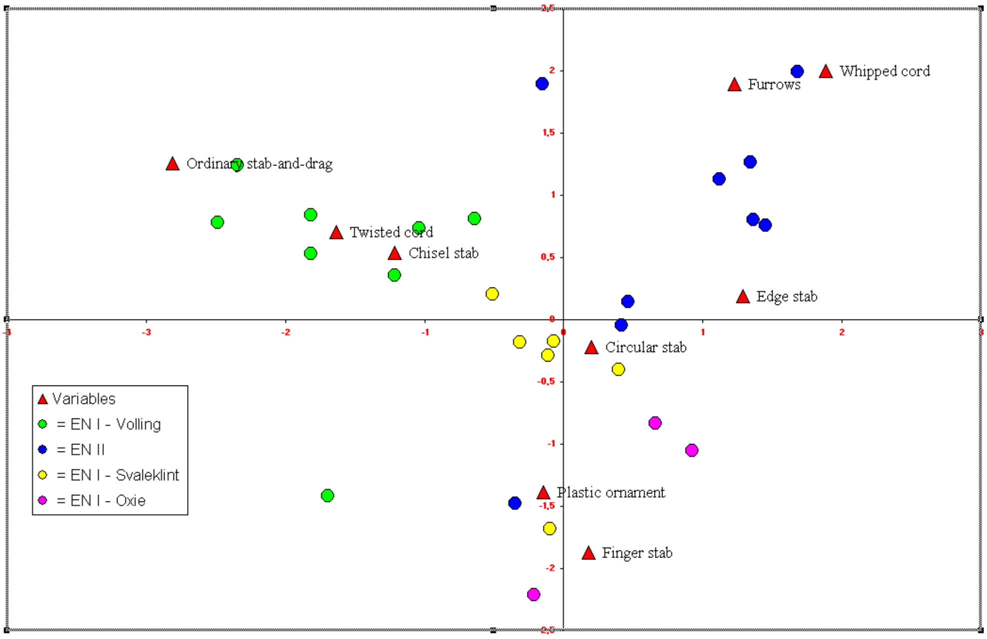
CA of the occurrence of ten different technical rim decoration elements in 27 settlements. Objects with sums larger than 100 are weighted down to 100. Variables are not weighted. Combined plot of 1. and 2. principal axis.

We can now return to the problem with outliers.



CA of the occurrence of ten different technical rim decoration elements in 27 settlements. Objects with sums larger than 100 are weighted down to 100. Variables are not weighted. Combined plot of 1. and 2. principal axis showing size of inertia.

If we create a plot showing the inertia of objects and variables we can locat the objects and variables that exert the greatest influence on the result. In the upper right hand corner there is one site (*Bistoft*) and one variable (*Furrows*) isolated from all other object and variables with very high inertia percentages. In the upper left hand corner there is a similar combination of an object (*Bønnerup*) and one variable (*Broad Stab-and-drag*) with high inertia percentages. An inspection of the input data reveals that 74% of the technical elements at *Bistoft* are *Furrows*, which by itself is a fairly common element on many sites. Further, 87% of all *Broad stab-and-drag* elements are found at *Bønnerup*, a site that contains a wide variety of other elements. Clearly, the site of *Bistoft*, and the variable *Broad stab-and-drag* are extreme outliers in the analysis and they should be removed. There are other variables with an inertia that should be dampened. In the following, apart from removing the site of *Bistoft* and the variable *Broad stab and drag*, the variables, *Whipped cord, Twisted cord*, *Ordinary stab-and-drag* and *Furrows* have been weighted with 0.7, 0.2, 0.7 and 0.8 respectively. This has been done experimentally to land their inertia on a not too high level compared to the other variables.



CA of the occurrence of ten different technical rim decoration elements in 27 settlements. Objects with sums larger than 100 are weighted down to 100. Variables have been individually weighted. Combined plot of 1. and 2. principal axis showing size of inertia.

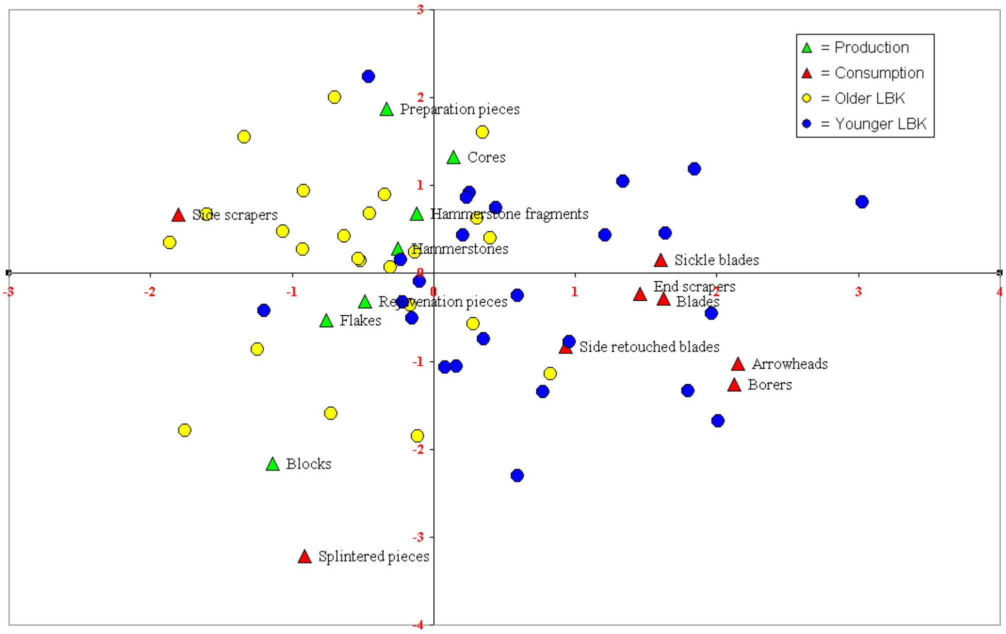
The exclusion of a site and a variable, and the selective weighting of variables have created a somewhat clearer result with a better separation of the individual groups. However, there are clearly problems with some sites (the three at the bottom centre of the plot), and after removing sites with a sum of less than 10 there are hardly any Oxie group sites left. Thus we may conclude that although the technical elements used in rim decorations appear to be a strong indicator for the chronological and cultural division of Early Neolithic pottery it is not by itself a sufficient discriminator.

Example of flint distribution in relation to Linear Band Ceramic houses

The material for this example is adopted from de Grooth 1987. It consists of counts of various types of worked flint from rubbish pits associated with houses from the Linear Band Ceramic culture (LBK) site of Elsloo. The material was expected to be able to reveal different potential modes of production. Basically *domestic mode of production* on the one hand and *Lineage mode of production* and/or *Loose mode of production* (*ad hoc* specialisation) on the other hand. The first would result in a uniform distribution of leftovers from production in all houses. The latter would result in a bipartition of houses into production and consumption units. As the different modes of production are not mutually exclusive a clear patterning cannot be expected, and only by using multivariate methods can patterns be uncovered (de Grooth 1987: 38).

A PCA was applied to the data, but no clear results were obtained, and none that related to the modes of production. In fact “rather unexpectedly, the only way to make sense of the two first P(rincipal)C(omponent)’s was to interpret them in chronological and technological terms. As time went on, fewer preparation pieces were needed to prepare cores that yielded a higher proportion of blades” (de Grooth 1987: 42).

PCA is not the correct method to use on contingency data, however. CA can be expected to perform much better, which will be demonstrated in the following. The material consists of counts of 19 categories of worked flint from 72 house units. The sum of counts for both flint categories and house units vary considerably. For the former the maximum sum is 3081, while the minimum sum is 1. For the latter the maximum sum is 762 and the minimum sum is 5. It was decided to leave out houses with a sum less than 20 and flint categories with a sum less than 15 to reduce the likelihood of a random effect. This leaves us with 49 houses and 15 categories of worked flint. The houses can be dated within a six phase chronology based on pottery. For this example it is sufficient, however, to split them into a group of houses dating to the older LBK and a group of hoses dating to the younger LBK. The 15 categories of worked flint used can be divided into a group of categories indicating production and a group of categories indicating consumption. The former group consists of *Blocks*, *Cores*, *Flakes*, *Rejuvenation pieces*, *Preparation Pieces*, *Hammer stones* and *Hammer stone fragments*. The latter group consists of *Blades*, *Side retouched blades*, *Sickle blades*, *End scrapers*, *Arrow heads*, *Borers*, *Side scrapers* and *Splintered pieces*.



CA of the occurrence of 15 categories of worked flint in 49 house units. Objects with sums larger than 100 are weighted down to 100. Variables are not weighted. Combined plot of 1. and 2. principal axis.

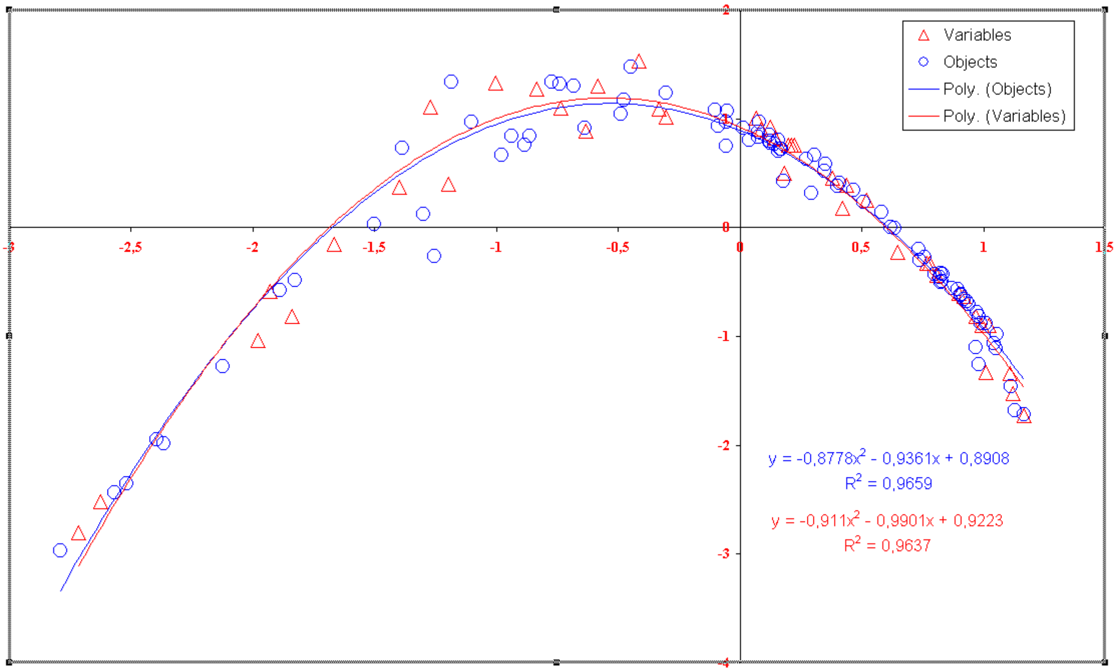
The result of the analysis is clear and directly interpretable along the lines that de Grooth had envisaged. There is a tight group of consumption categories in the right part of the plot and a somewhat more dispersed group of production categories to the left and around the centre of the plot. This bipartition clearly indicate that different modes of production are indeed imbedded in the material. Only two assumed consumption categories break the pattern by lying to the left and bottom of the plot. *Splintered pieces* may not belong to the consumption category as I have assumed, and *Side scrapers* seems to be a tool that mostly belongs to the older LBK, and hence may be caught up in the chronological pattern.

The houses have an even distribution with most of them lying around the production categories, and considerably fewer around the consumption categories. The really surprising fact is that all houses from the older LBK lies around the production categories, while the houses from the younger LBK are evenly distributed around both groups. The interpretation seems fairly clear.

During the older LBK the domestic mode of production prevailed, and if other modes existed they were not sufficiently developed to pattern the material. During the younger LBK other modes of production broke through (whether lineage mode or loose mode), and house holds are now either producers and consumers or primarily consumers.

Example of female graves from the Germanic Iron Age on the isle of Bornholm

The data for this example is adopted from Nielsen 1988, which was one of the very first studies showing the capabilities of CA for chronological studies, demonstrating how, in connection with perfect continuity in a set of data, the graphical presentation would form an arced hyperbolic layout. Here the analysis is merely presented with a few comments.



The data consist of counts of various types of personal ornaments in female graves. As can be seen from the plot there is a high degree of continuity in the material, where the individual ornament types occurs in a fairly limited number of graves, and where each grave have a limited number of ornaments. Further, there are no breaks in the sequence leaving us with a perfect seriation that can be interpreted chronologically. The tendency for clustering along the hyperbolic layout may either indicate an uneven temporal occurrence of graves in the material, or it may be the result of an uneven temporal development in the type of ornaments used.

Clearly, the layout can be described through a second degree polynomial. In the above plot separate trend lines for objects and variables have been added. In CAPCA this can easily be done in the plots, when objects and variables are shown without a classification. Just activate the series of object, right click and choose add trend line and make sure that the type is set to polynomial. Then repeat the process with the variables. When adding the trend line you can also specify to have its equation shown as well as the squared value of Pearson’s correlation coefficient. For a good seriation you should expect this value to be very high. No rules can be give, but I would expect it always to be higher than 0.9. Further in a good seriation you should expect the two trend lines to be almost identical.

To most archaeologists a seriation is equivalent to an ordered matrix, where values are concentrated along the diagonal. Such a sorting can of course be done from the CA result. In CAPCA a sorted matrix is always shown on the sheet *Matrix output* when you run a CA. This is irrespective of whether the CA produce a seriation or not, and you should always decide from the CA plots, whether you have a seriation or not, before you turn to the sorted matrix. *Never* judge a seriation from the sorted matrix!

	nv	m3	r3d	n5	q3c	n3	q3b	r3c	n4	o4	o3	i3	n1+2	q3a	k1a	i2	r3b	k1c	s1	q3d	k1b	s3	q2	i1d	d	p4	p2	g3	e2a	e2b	p3	r3a	r1	g2	i1b	g1	p1	h	f	i1a	a2e	e1
Lillevang 9	2	1																																								
Lousgård 2	2	1																																								
Lousgård 2	2		1																																							
Lillevang-M	2				1																																					
Nr.Sandeg	2				1																																					
Kobbeå 20		1	1																																							
Lillevang 1	1				1																																					
Bækkegård 15		1			1	1																																				
Lousgård 47			1	2	1		1																																			
Saltuna 14			1	2	2		2																																			
Lillevang-Melsted			1				1		2																																	
Lillevang-Melsted 4						1	1	1		1																																
Bækkegård 105					1	1			1								1																									
Lillevang 2			1		1				1													1																				
Gudhjem								1		1																																
Ellegård							2	1			2		1																													
Lousgård 12					1		1				1		2				1																									
Bækkegård 153							1		2			1		1			1																									
Bækkegård 132							1										1																									
Bækkegård 44					1						2					1	1																									
Bækkegård 59										1	2						1																									
Bækkegård 66										1	1						1																									
Lousgård 6										1	1						1																									
Lousgård 11											2						1																									
Lousgård 3											2						1																									
Nr.Sandegård 6											2						1																									
Bækkegård 143									1							1	1																									
St.Kannikegård 195													1				1																									
Bækkegård 3												1				1	1																									
Bækkegård 5														1	2		1																									
Bækkegård 77														1	1	2	1								1																	
Bækkegård 50																1	1			2																						
Bækkegård 76																1	1									1																

The sorting of objects and variables in CAPCA is based on the coordinates of the first principal axis. This is common practice, and presently the only viable approach to an automated sorting. It is not the optimal solution. The sorting order should follow the polynomial trend line rather than the first axis. For a good seriation this would yield a more or less identical result with the one obtained from the first axis, but where objects and variables are more widely distributed on both sides of the trend line, the sorting order can vary considerably. Automated sorting by way of the trend line, however, awaits a suitable algorithm.

METRIC SCALING

Measures of similarity and distance

The development and use of similarity and distance coefficients took place in the biological and ecological sciences in the 1950'es and 60'es. A central publication was "Principles of Numerical Taxonomy" (Sokal & Sneath 1963), which also had an impact on archaeology through David Clarkes "Analytical Archaeology" (1968).

The concept of similarity is one we all share, but it is certainly not a precise and well defined concept. In every day life we often state that something is similar or dissimilar, but if questioned as to why we will often find it difficult to give a precise answer. The problem is that we cannot really speak of similarity unless we also state in terms of what. Even when we intuitively speak of similarity between objects we do so based on abstractions from the objects of characteristics that we feel are important. We may not realise what these characteristics are, and if questioned we may be at a loss to explain. Others may not find the same similarity because they focus on other characteristics.

To use the concept of similarity in science we have to define what similarity is and how we measure it. It is agreed that a statement of similarity between objects is based on a predefined list of elements/traits/characters – whatever the name used – through which the comparison is made. It is obvious that similarity becomes a matter of the predefined list. In the heydays of positivism this indicated that if only we could set up a thorough list of elements we could reach an objective statement of similarity. This is obviously not so. Many different lists of elements can be set up reflecting current understanding and goals. Each will result in different statements of similarity. However, given a particular list of elements and given a particular way of measuring similarity based on this list we can get consistent and repeatable expressions of similarity between objects.

A variety of measures of similarity have been suggested and used over the years. Most measures result in coefficients of similarity ranging between 1 and 0, the former for perfect agreement, the latter for no agreement whatever. The measure used in CAPCA is adopted from J.C. Gower (1971). This is a generally approved measure that elegantly combines elements from three types of variables. The three types separated are: dichotomous, qualitative and quantitative. A dichotomous variable holds one element which must be either absent or present for an object. A qualitative variable has two or more alternative elements. Only one element can be recorded for an object and the object must always display one of the alternatives. A quantitative variable has a set of numeric values with an inherit order. It may be measurements, counts or even numbers representing an ordinal scale.

When comparing two objects across all their variables two “counters” are used called *Scores* and *Validity*. Whenever a valid comparison between two variables is made *Validity* is incremented with 1, while *Scores* is incremented with a value between 0 and 1 depending on the outcome of the comparison.

For dichotomous variables *Scores* is incremented with 1 if both objects show presence and is not incremented if one object shows presence and the other shows absence. If both objects show absence the comparison is not seen as valid and neither *Scores* nor *Validity* are incremented. For qualitative variables *Scores* is incremented with 1 if both objects display the same element and is not incremented if they differ. For qualitative variables *Scores* is incremented with a value calculated as $1-|x_i – x_j|/r$ where x_i and x_j represent the values of the variable for the two objects and r denotes the total range of values in the variable.

Gowers coefficient is the only coefficient supported in CAPCA if the input data consist of objects and variables as recorded. However, it is also possible to input matrices of coefficients directly, but then you have to compute the coefficients yourself in advance. These matrices of coefficients may contain either similarity coefficients or distance coefficients, the latter in principle being just a reciprocal expression of similarity.

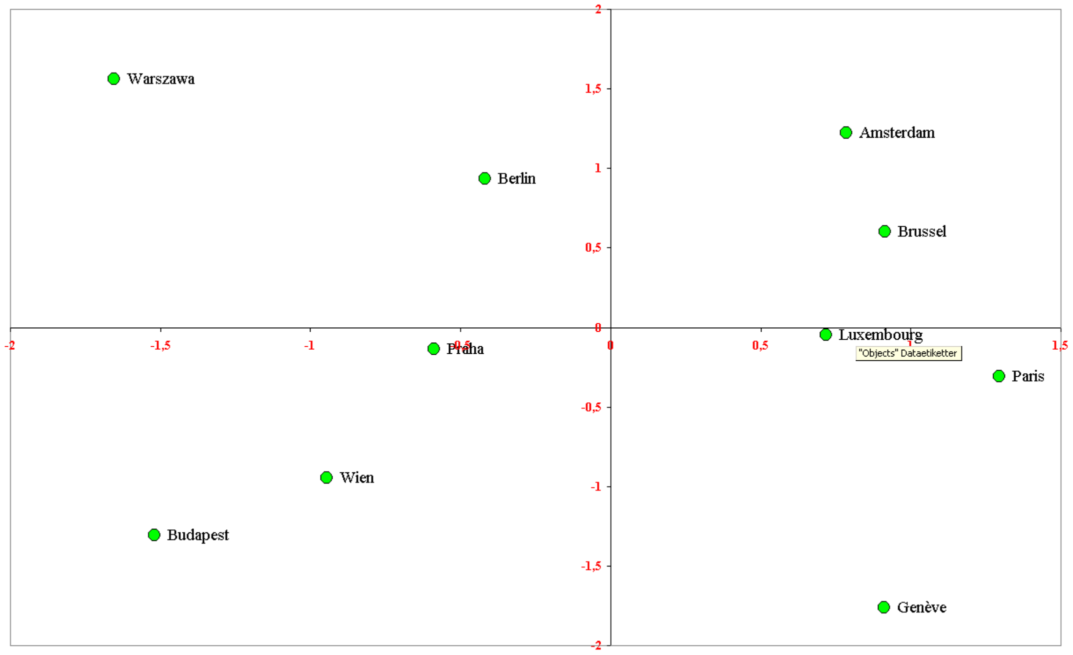
Example of distances between Central European capitals

In connection with most roadmaps you will find a table of distances between cities. The following example was drawn from a standard road atlas of Europe. Ten capitals have been extracted from the table, all of them lying within mainland Central Europe. Thus you don’t have to cross the sea to go from on capital to another, and none of the capitals lies on peninsulas restricting the driving access.

	Amsterdam	Berlin	Brussel	Budapest	Genève	Luxembourg	Paris	Praha	Warszawa	Wien
Amsterdam	0	668	211	1411	908	383	501	855	1226	1152
Berlin	668	0	777	859	1079	766	1052	343	595	625
Brussel	211	777	0	1367	714	215	309	891	1335	1108
Budapest	1411	859	1367	0	1284	1190	1494	517	669	247
Genève	908	1079	714	1284	0	508	503	922	1559	1025
Luxembourg	383	766	215	1190	508	0	355	725	1287	930
Paris	501	1052	309	1494	503	355	0	1030	1611	1234
Praha	855	343	891	517	922	725	1030	0	614	283
Warszawa	1226	595	1335	669	1559	1287	1611	614	0	695
Wien	1152	625	1108	247	1025	930	1234	283	695	0

Distance matrix between 10 Central European capitals. The distances are kilometres.

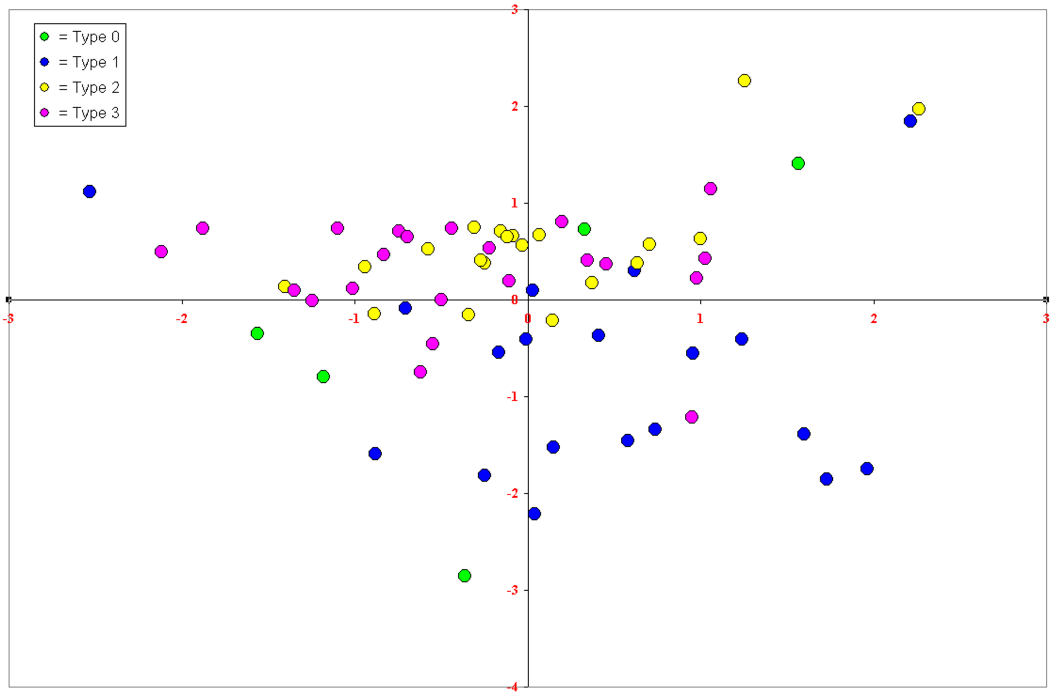
The input matrix is a straight forward distance matrix with zero values on the diagonal and distances in km in the cells combining various capitals.



Metric scaling of distance matrix between 10 Central European capitals. Note that the “map” is mirrored.

The result of the metric scaling of the distance matrix is a fairly precise representation of the position of the capitals in relation to each other. You can reassure yourself of this from any map of Europe, but you have to do a little mental mirroring of the map. East has become west and vice versa. Obviously there is no way that the program can know what is left and right or up and down. It is simply a scaling presented in two dimensions from some distance measures. If you want to compare it to the real world, you have to do a bit of mirroring yourself, and perhaps even rotating.

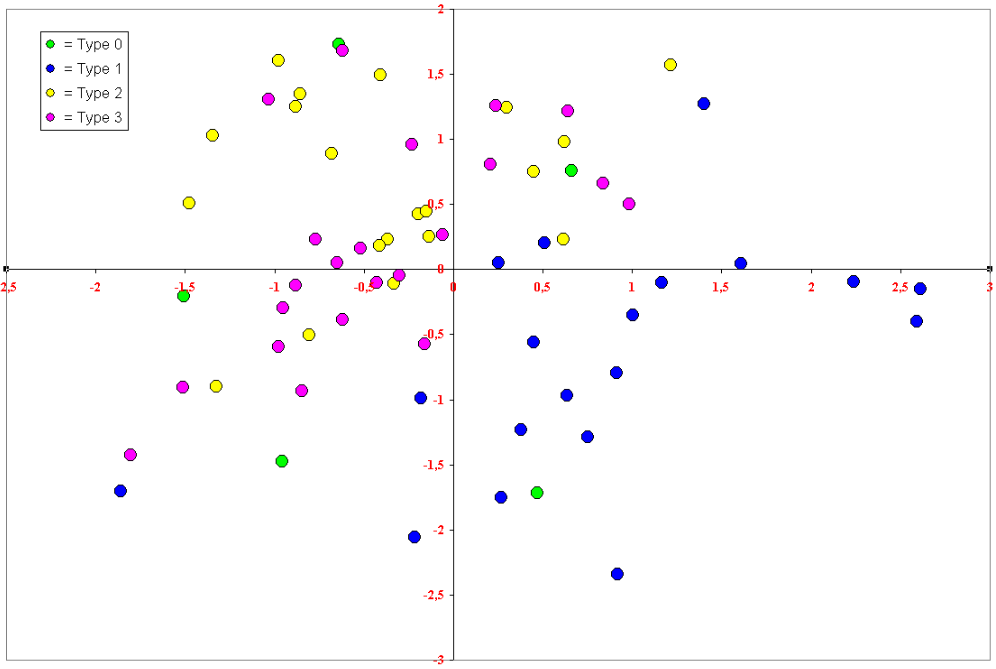
Example using measurement and rim decoration data on 66 early Neolithic pots
This material is the same as was used in a PCA example above, but in addition to the measurements used, a recording of decoration elements is added by way of Gowers general coefficient of similarity. First, however, an analysis exclusively based on the measurement data is made to enable a comparison between PCA and MS.



Metric scaling of 66 Neolithic pots based on a comparison of 12 measurements.

The measurements used are in the weighted version that brought the best results with the PCA. It can immediately be seen that the result of the MS is markedly inferior to the result from PCA. There is a tendency for a separation of type 1 on the one hand and type 2 and 3 on the other, but that is just about all. In the PCA we actually had a clear separation of type 0 and type 1 on the one hand and type 2 and type 3 on the other, and a partial separation of the two latter types, all fully in line with what we would expect archaeologically. Why things do not work out in MS is difficult to evaluate, not least because we totally loose the connection to the individual variables in the process of creating the similarity coefficients. In my opinion, however, a coefficient of similarity is too simple a way to express the relations between objects.

The main reason to use MS is the possibility to combine continuous variables with categorical variables. Unfortunately, the pots in question are not highly decorated, in fact the majority have only a rim decoration, if any decoration at all. Further, the rim decoration is quite simple, displaying only horizontal lines or rows. Basically, we are limited to record 11 different technical elements used in the rim decoration. Technical elements, however, can be considered to be very decisive in the Early Neolithic pottery as one of the CA examples above shows.



Metric scaling of 66 Neolithic pots based on a comparison of 12 measurements and 11 technical elements in rim decoration.

The MS of measurement and decoration data together is not satisfactory either. Clearly, there is an improvement in the separation of type 1 from the others, but there are no significant changes within the remaining material. The use of MS in this case is simply not satisfactory. It remains to be seen whether this is a reflection of a general weakness in the approach compared to PCA/CA, or if there are situations, where better results can be obtained by MS than by combining your way with PCA and CA. Personally, I am rather sceptical.

LITERATURE

Baxter, M.J. 1994 *Explorative Multivariate analysis in Archaeology*. Edinburgh University Press, Edinburgh.
Clarke, David L. 1968 *Analytical Archaeology*. Methuen & Co Ltd, London.
de Grooth, M.E.Th. The Organisation of Flint Tool Manufacture in the Dutch Bandkeramik. *Analecta Praehistorica Leidensia 20*, 1987, p. 27-51.
Gower, J.C. 1971 A general coefficient of similarity and some of its properties. *Biometrics 27*, p. 857-74.
Jensen, Claus Kjeld & Karen Høilund Nielsen 1997 Burial Data and Correspondence Analysis. In Claus Kjeld Jensen & Karen Høilund Nielsen (eds.) Burial & Society. The Chronological and Social Analysis of Archaeological Burial Data. Aarhus University Press, p.29-61.
Nielsen, Karen Høilund 1988 Correspondence analysis applied to hoards and graves of the Germanic Iron Age. In Torsten Madsen (ed.) *Multivariate Archaeology. Numerical Approaches in Scandinavian Archaeology*. Jutland Archaeological Society Publications XXI, Aarhus University Press, p. 37-54.
Shennan, Stephen 1988 *Quantifying Archaeolo*gy. Edinburgh University Press, Edinburgh.
Wright, R. 1985 Detecting pattern in tabled archaeological data by principal components and correspondence analysis: programs in BASIC for portable microcomputers. *Science and Archaeology 27*, 35-38.
Sokal, Robert R. & Peter H.A. Sneath 1963 *Principles of Numerical Taxonomy*. W.H. Freeman and Company, San Francisco.