

# CAPCA Version 2.2

## Torsten Madsen ©2012

### INTRODUCTION

CAPCA is an add-in to Microsoft Excel that allows you to carry out Principal Components Analysis (PCA), Correspondence Analysis (CA) and Metric Scaling (MS) within and fully integrated with the Excel environment. Filling in references to data ranges and selecting various options in a form is all it takes to run the analysis produce a worksheet with statistics showing the result of the analysis and create a number of charts to illustrate it.

For further information on PCA, CA or MS and various examples on their application to archaeological data you should read the paper “Multivariate data analysis using PCA, CA and MS CAPCA”.

### INSTALLATION

Installing CAPCA on your computer is simple. The only precondition is that Microsoft Excel (Office XP or later) is installed.

*Office XP or Office 2003*

You copy the CAPCA.xla file to the directory *AppData/(Roaming)/Microsoft/addIns*. You will find this as a sub directory to the directory for your personal settings (e.g. *Users/user name*). (NB! The directory may be hidden. In that case you need to change the settings in the file browser allowing you to see hidden directories).

In the *Tools* drop down menu in Excel you choose *AddIns*. In the subsequent AddIns dialog box you check the *CA - PCA analysis* option, which will add a new item to the Tools menu named *CA - PCA analysis*. You click this option to display the CAPCA input form.

*Office 2007*

You copy the CAPCA.xlam file to the directory *AppData/(Roaming)/Microsoft/addIns*. You will find this as a sub directory to the directory for your personal settings (e.g. *Users/user name*). (NB! The directory may be hidden. In that case you need to change the settings in the file browser allowing you to see hidden directories).

Press the “office button” and select Excel options at the bottom. In the menu that appears you select AddIns. In the subsequent AddIns dialog box you check the *CA - PCA analysis* option, which will add a new item to the *AddIns* menu named *CA - PCA analysis*.

If CAPCA is already present here (via CAPCA.xla) you disable it. Next you remove CAPCA.xla from the addins library. Finally, you enable CAPCA again in the addins dialog box. Probably, you will be told that the file does not exist and is offered the choice of browsing for it. Do this and choose CAPCA.xlam.

### DATA INPUT

All data to be analysed and all auxiliary information must be placed in an Excel worksheet. The input form displayed, when you activate CAPCA, is only used to provide the program with references to the position of the different sets of information. Such references must follow the standard notation rules for referring to cells in Microsoft Excel. Thus A1:T36 refers to the rectangular block of data placed between cell A1 in the upper left corner and cell T36 in the lower right corner. It is here assumed that the data is in the current worksheet - that is the worksheet visible behind the input form. ***Running an analysis where some or all of the data are not present in the current worksheet will result in run time errors!*** If you wish to provide data that are not in the current worksheet you must use extended references (e.g. Sheet1!A1:T36). In this way you may

also provide data from different worksheets for one analysis (e.g. Values in one worksheet and names for variables and objects in another).

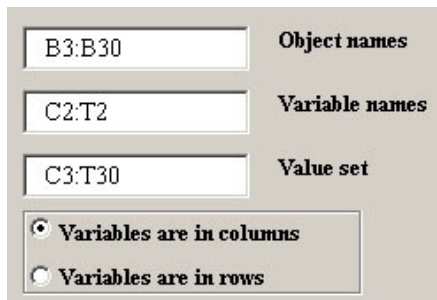
There are some data that you must provide in order to run PCA, CA or MS and others that you can provide optionally.

The mandatory data needed to run an analysis are: a matrix of values to be analysed; an array of object names; an array of variable names.

The optional data includes: an array of object class names; an array of variable class names; an array of object weights; an array of variable weights.

### Mandatory data

The input form contains three boxes for references to the position of mandatory data named *Object names*, *Variable names* and *Value set*. You must provide references to cell ranges in all three for PCA, CA and MS. However, if for MS you provide a matrix of coefficients instead of a matrix of objects and variables, the reference to variable names is not needed and the input box will be disabled.



B3:B30	Object names
C2:T2	Variable names
C3:T30	Value set
<input checked="" type="radio"/> Variables are in columns	
<input type="radio"/> Variables are in rows	

The reference to a cell range for object names must point to a one dimensional array. That is, all object names must be placed in either one continuous column (e.g. B3:B30) or one continuous row (e.g. B1:AK1).

The reference to a cell range for variable names must point to a one dimensional array. That is, all variable names must be placed in either one continuous row (e.g. C2:T2) or one continuous column (e.g. B3:B20).

The reference to a cell range for the value set must point to a two dimensional array (matrix). That is, the values must be placed in a continuous rectangular block (e.g. B1:T36 or if variables are in the rows C3:AD20). The variables may be placed in the columns or they may be placed in the rows, but as the program must know which is which there is an option box below the reference to the value set, where it is indicated whether *Variables are in the columns* (default) or *Variables are in the rows*.

It is common practice and a very good idea to place the variables in the columns and the objects in the rows. Further, the row immediately above the value set should contain the variable names as column headers and the column immediately to the left of the value set should contain the object names as row headers. In this way names and values always follow each other in a standardised manner.

An object or a variable name can be any string of text up to 255 characters. You should, however, always choose as short and precise names as possible. Long and complex names will clutter the output. Especially the graphs will suffer from too much textual information.

The value set must consist of digits and nothing else. No letters are allowed. Blanks are treated differently in connection with PCA, CA or MS. In typical data for PCA (e.g. measures of some kind) missing values are fairly common and is often represented by blanks. As missing values are not allowed in a PCA blanks are here considered as illegal values. In a CA missing data is normally not an issue (still they are not allowed), but often you have huge data sets where more than 90% of the cells contain zeroes. Hence, blanks are here treated as zeros allowing you to create value sets without filling in all the zeroes. Further, it is also an advantage not to have the zeros in the worksheet, as they make it more difficult to audit the recordings. In MS blanks are treated as illegal values for variables declared to be continuous and as zeros for variables declared to be categorical.

For PCA any real number is acceptable as the values are expected to be some kinds of measures. In CA only positive integers are acceptable as the values are expected to derive from contingency

tables. In MS the acceptable values depend on the declaration of the variables. For continuous variables any real number is acceptable, whereas for categorical variables only 1 and 0 is accepted.

### Optional data

For PCA and CA it is possible to associate a classification with both objects and variables, while for MS it is possible to associate a classification with objects only. It can be anything that divides the objects or variables into sets: a regular classification; a chronological scheme; a geographical division. For each object or variable you simply provide a class name, and the number of different class names you provide automatically becomes the number of classes that will appear in plots later on in the analysis.

Setting a checkmark in the checkbox *Add object classification* will display an input box named *Object class names*. You must provide a reference to a range of cells in this box, whenever it is visible. The cell range reference must point to a one dimensional array. That is, all object names must be placed in one continuous column (e.g. A3:A30) or one continuous row (e.g. C1:AD1). Preferably

you should place the object classification in a column to the left of the object names.

Setting a checkmark in the checkbox *Add variable classification* will display an input box named *Variable class names*. You must provide a reference to a range of cells in this box, whenever it is visible. The cell range reference must point to a one dimensional array. That is, all object names must be placed in one continuous row (e.g. C1:T1) or one continuous column (e.g. A3:A20). Preferably you should place the variable classification in a row above the variable names.

For PCA, CA and MS, it is possible to associate weights with both objects and variables. For MS, however, only in a very restricted manner, as you can only provide weights if your data are in the form of objects described by variables, and not if they are in the form of a coefficient matrix. Further, the only weights acknowledged are 1 and 0 (for inclusion or exclusion – see below for further information on weights), as a proper weighting factor would have catastrophically results if applied to values of categorical variables.

At the right hand top of the input form there are two checkboxes named *Apply weights to objects* and *Apply weights to variables*.

Setting a checkmark in the checkbox *Apply weights to objects* will display an input box named *Object weights*. You must provide a reference to a range of cells in this box, whenever it is visible. The cell range reference must point to a one dimensional array. That is all weights must be placed in one continuous column (e.g. V1:V30) or one continuous row (e.g. C22:AD22).

Setting a checkmark in the checkbox *Apply weights to variables* will display an input box named *Variable weights*. You must provide a reference to a range of cells in this box, whenever it is visible. The cell range reference must point to a one dimensional array. That is all weights must be placed in one continuous row (e.g. C32:T32) or one continuous column (e.g. AF3:AF20).

NB! If you use weights for either objects or variables you must have a weight for each and every variable or object. You cannot leave any part of the range in your reference empty.

It is common practice and a very good idea to place the variable weights in a row immediately below the values of the data set and the object weights in a column immediately to the right of the

values in the data set. The complete set of common practices regarding data layout is shown below (top of page 4).

Using weights in PCA and CA may perhaps seem controversial, but as conceived of here the use of weights is more than just a question of adjusting the influence of individual variables or objects in the analysis. It is also a mean to control which variables or objects should enter the analysis.

Since a weight must be provided for every variable/object the first thing we need is a neutral weight. This is set to 1. If you fill in the complete range of weights with 1 no changes what so ever will occur to your analysis.

		Object Classification																Variable Names				
		Object Names																Variable Classification				
1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
2			V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18		
3	a	U1			1		1	1	1	1		1	1		1							1
4	a	U2	1					1									1	1				1
5	a	U3		1						1			1		1						1	1
6	b	U4			1		1										1			1		1
7	b	U5				1		1	1													0
8	b	U6		1						1	1	1	1									1
9	a	U7	1		1			1		1					1	1	1	1				1
10	c	U8										1							1			0
11	c	U9			1												1					0
12	c	U10				1	1	1	1			1			1						1	1
13	c	U11								1						1	1		1			1
14	b	U12	1	1	1			1								1	1			1		1
15	b	U13														1		1			1	1
16	a	U14			1		1		1									1	1			1
17	c	U15								1						1						1
18	c	U16		1				1					1	1				1	1			1
19	c	U17				1			1	1					1						1	1
20	a	U18			1					1							1		1			1
21	a	U19	1				1		1							1						1
22	b	U20						1	1				1									0
23	b	U21		1						1	1			1	1	1	1	1	1			1
24	b	U22			1	1														1		0
25	a	U23					1		1			1		1			1					1
26	c	U24	1				1					1	1								1	1
27	c	U25		1												1	1		1			1
28	b	U26			1		1	1	1	1	1									1		1
29	b	U27									1				1		1					0
30	a	U28			1		1			1			1					1	1			1
31																						
32			0	1	1	0,5	1	1	0	0,8	0	1	1	1	1	0,9	1	1	1	1	0	

The second thing we need is a weight that will remove a variable or an object from the analysis. This is set to 0. Especially in CA based on presence/absence recordings it is very often necessary to remove both objects and variables that tend to form unique structures not co-varying with the majority of objects and variables. Using weights of 0 is an elegant way of doing this in stead of having to edit the initial table of input data.

Values between 1 and 0 will be used as a factor that is multiplied with the values of the object or variable in question. Allowing values between 1 and 0 only means that you can reduce the influence/significance of an object or a variable, but you cannot increase it.

Bear in mind that the weights are values in cells of a spreadsheet. The direct approach is of course to write the values directly into the cells. However, you may also write functions in the cells letting the spreadsheet calculate weights based on the values in your data set. Thus, in connection

with a CA you may for instance write a function that will substitute the absolute values of a variable with percentages if the variable sums up to more than 100.

### Number of eigenvectors and precision of calculation

You can specify the *Number of eigenvectors to be calculated* (Principal components or Principal axes) in the appropriate input box. The default number is 3, but you can set the value to anything between 2 and the maximum number of eigenvectors possible. If you write too large a number it will automatically be reduced to the maximum. If you write too small a number (1, 0 or a negative number) the number will default to three. The same applies if what you write cannot be interpreted as a number.

The calculation of Eigenvectors in CAPCA is based on an iterative algorithm published by Wright (1985). The level of precision of the calculation is determined by a control number that is calculated after each loop. When this number reaches a preset level (becomes small enough) the analysis concludes. It is important here to know a little about how the iterative algorithm behaves.

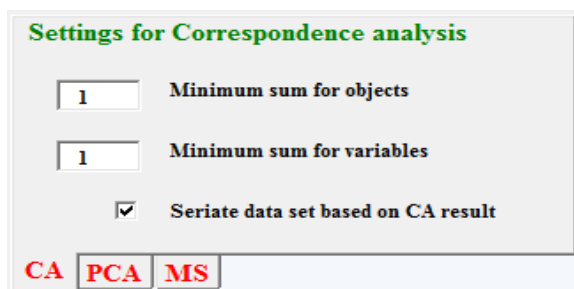
As the algorithm goes through its iterative loops of calculations the results it obtains on a loop will become more and more identical with the results of the previous loop, and hence more and more accurate. However, this growing stability of results does not happen equally across all calculated axes. Stability is reached progressively through the sequence of principal axes, and indeed it is questionable if stability can ever be reached for the last principal axes in a large CA, as convergence proceeds towards infinity.

If we set a low starting value for the control number the analysis will terminate very quickly, but only the first few axes calculated will be stable. If we set a higher starting value for the control number the analysis will take longer to complete, but a larger number of the axes calculated will be stable. In version 2 the starting value is permanently set very high to make certain that instability cannot occur on the first few axes. This is in contrast to version 1 where the user could control the value, leading in some proven cases, when low precision was chosen, to instability on the second principal axis. In version 2 you can always depend on absolute stability for the first three axes, and under normal circumstances there should be stability up to ten axes. Above ten you cannot expect stability.

### Optional settings

*Setting the minimum sum of objects and variables for CA*

For CA two input boxes named *Minimum sum of objects* and *Minimum sum of variables* is available. The two input boxes will both be shown with a default value of 1. You can fill in any number larger than 1.



Settings for Correspondence analysis

Minimum sum for objects

Minimum sum for variables

Seriate data set based on CA result

CA PCA MS

In a CA we will very often have to deal with very small sums, especially for objects, but occasionally also for variables. With presence/absence data (i.e. the content of graves), we may have situations where an object holds only one item (a grave contains one artefact). It is not meaningless to include this grave in the analysis, but depending on the kind of results we seek, it may be more sensible to analyse only those objects that shows combinations of items

(those graves that have at least two artefacts), and we would thus set the minimum sum of objects to two.

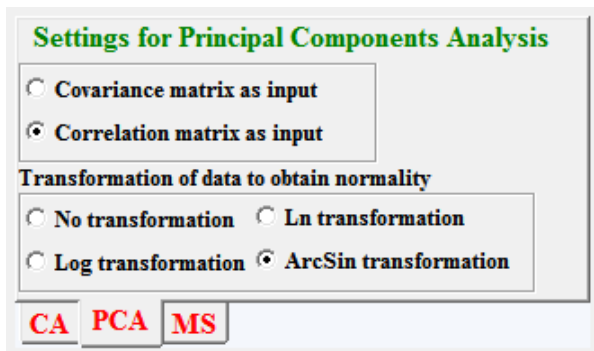
With abundance data (i.e. number of items of various types in a settlement) we may be concerned with how representative the counts are. It is well known that we should not calculate

percentages with sums that are lower than 50 (ideally not lower than 100). CA is based on relative distributions, and although for various reasons the effect of small samples is not as devastating as with percentages, there are good reasons to set a minimum sum for the objects. How low you should set the minimum is an archaeological decision. If you expect your objects to be randomly constituted (i.e. debris) you may choose to set a relatively high minimum value. If you expect your samples to be meaningfully constituted through deliberate choices in the past (i.e. content of hoards) you may choose to set a low minimum value.

When you run CA, the program will check for any sum of objects and variables smaller than the ones stated as minimum, and exclude those found from the analysis. Exclusion means that sums of objects and/or variables are altered. Therefore the exclusion procedure will run repeatedly until no more exclusion and hence no more alterations in sums occur.

Checking the box *Seriate data based on CA result* will add a worksheet named *Seriation output* containing a matrix sorted according to the equation for second order polynomials through the objects and variable respectively on the two first principal axes. **Please note:** do not seriate data unless you have decided through an inspection of the graphical output that a seriation is meaningful.

### Choosing between a covariance and a correlation matrix for PCA



For PCA you can decide what kind of standardisation of data should take place before analysis. There are two options

Covariance matrix:

For each value of a variable we subtract the mean value of the variable and divide by the square root of the total number of values minus one.

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

In this way PCA will be based on what is known as the covariance matrix.

Correlation matrix:

For each value of a variable we subtract the mean value of the variable and divide by the standard deviation of the variable multiplied with the square root of the total number of values minus one.

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

In this way PCA will be based on what is known as the correlation matrix. This is the default setting.

### Choosing scale transformations of data for a PCA to obtain normality

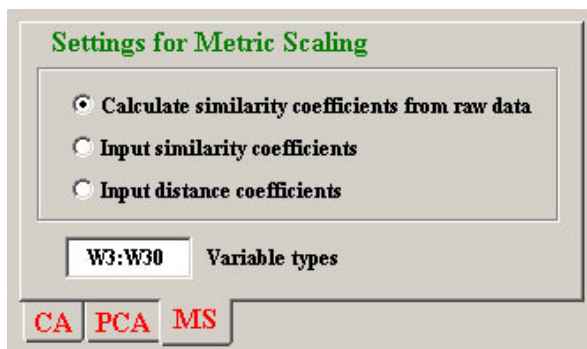
If you are using a correlation matrix as input, then ideally all variables should be normally distributed. When dealing with measurement data this is seldom the case. In most cases the distribution will be skewed towards the higher values.

One way to counter this problem is to change the scale of the variables through some form of numerical transformation. The most common transformation to use is a logarithmic transformation. In CAPCA, at the moment, transformations based on either  $\text{Log}(10)$ ,  $\text{Ln}$  (the natural logarithm) or  $\text{Arc Sin}$  are implemented.

The default setting for this option group is *No transformation*. Alternatively you can choose *Log transformation*, *Ln transformation* or *ArcSin transformation*. Log and Ln transformations will only work with positive data. If your data contain negative values and you have chosen either a Log or a Ln transformation you will get an error message during the screening process of data. You will not be able to choose transformations, if you have chosen the covariance matrix input.

### Choosing the type of input data for Metric scaling

Metric scaling operates on either similarity coefficients or distance coefficients. There are three different ways you can input data into CAPCA for metric scaling. You can either provide raw data in the form of objects and variables to let CAPCA calculate the similarity coefficients, or you can provide a matrix of either similarity coefficients or distance coefficients directly.



The default way to input data for a MS is as raw data, based on which the program will calculate the similarity coefficients. This is not without its problems. First of all the program will use one particular method of calculating the similarity coefficients. If you do not approve of this method you will have to produce the matrix of similarity coefficients yourself for direct input. Secondly, in order to calculate the similarity coefficients the program has to know a little about your variables.

The method used to calculate the coefficients is a slightly cut down version of Gowers general coefficient of similarity (Gower 1971). It is imperative here to distinguish between continuous variables and categorical variables. Continuous variables are all kinds of measurements, percentages or counts, although it would not be advisable to use raw counts. Categorical data in this situation are presence absence data and nothing else. In the paper *Multivariate data analysis with PCA, CA and MS* you can learn exactly how Gowers coefficient is computed.

You provide information on the variables type in a one dimensional array in a row or a column in the spreadsheet with the same size and sequence as the array containing the variable names. In the cells you write either *con* or *cat* for continuous and categorical respectively to identify the type of the individual variables. You then provide a reference to the row or column containing the type information in the textbox marked *Variable types*.

For the continuous variables you can provide any numerical value. Blanks are not allowed. For the categorical variables you can use 1 and 0 only representing presence and absence. Blanks are here interpreted as 0.

If you provide a similarity matrix directly, the following rules apply: all values must be positive; the highest possible values must be in the diagonal and all diagonal values must be the same (normal practise is either 1 or 100 as highest values); the values in the cells on one side of the diagonal must be a mirror of the values on the other side; the diagonal is here always defined as the

one leading from the upper left hand corner to the lower right hand corner. If you provide a similarity matrix with diagonal values higher than 1, this will be rescaled by CAPCA to a similarity matrix with diagonal values of 1 before analysis.

If you provide a distance matrix directly, the following rules apply: all values must be positive; all values on the diagonal must be zero; the values in the cells on one side of the diagonal must be a mirror of the values on the other side; the diagonal is here always defined as the one leading from the upper left hand corner to the lower right hand corner. The distance matrix will be transformed by CAPCA to a similarity matrix with diagonal values of 1 before analysis.

## **RUNNING THE ANALYSIS**

There is a multipage control to the lower right of the input form. This control has three pages marked CA, PCA and MS. The three pages hold information specific to the three types of analysis. By selecting one of these pages you also select the type of analysis to run. To start the analysis you press the run button to the lower left of the input form. If you have chosen the page marked CA the button will display *Run CA*. If you have chosen the page marked PCA the button will display *Run PCA*. If you have chosen the page marked MS this button will display *Run MS*. You should thus not be in doubt as to the type of analysis you are running.

When you start running an analysis a sequence of checks are performed on input data to ensure that the analysis will run as expected and without fault. At each step in the sequence of checks the analysis may be interrupted and an error message issued that informs you about the nature of the problem.

The sequence of checks each of which immediately terminates the analysis is as follows:

- 1) A syntax check on the range specification you have provided is carried out. If one or more of the range specifications do not meet the requirements of Microsoft Excel the following error message appears "The following are not proper Excel input ranges:" followed by the offending input ranges.
- 2) At this point you may get the error message "Unexpected information was found. The current worksheet may not be the sheet with your data". This is triggered during an attempt to read the data ranges you have provided. For instance, if a range exclusively expected to consist of numbers contain text, Excel will immediately issue the error code 438 (Object doesn't support this property or method). The most likely cause for the error is that the sheet with your data is not the current sheet (the one visible behind the input box), when you run the analysis, but it could also be due to your data containing letters where only numbers are expected. For instance an l instead of a 1 or an o instead of a 0. Looking for text in your data, you should focus on whether some of the cells are left justified in stead of right justified. As this error is triggered during reading of data it may appear at this point or later in the checking sequence.
- 3) The program checks if the references to Object names, Variable names, Object classifications, Variable classifications, Object weights, Variable weights and with MS Variable types are all pointing to a one-dimensional array. If any of these references include more than one row or one column, an appropriate error message is displayed.
- 4) The program uses the length of the object names array and variable names array to establish the number of objects and variables respectively. If either of these are less than three an error message is issued stating "There must be at least three rows and three columns of data". Please note that it may not be your data that are in error. It can also be your range reference to the column and row containing the names.
- 5) The program checks if the dimensions of the referenced ranges are in accordance with each other. Thus object names, -class and -weight must have the same length; variable names, -



class, -weight and for MS -types must have the same length; the dimensions of the data matrix must be in accordance with the number of objects and variables as defined by the object names and variable names. If errors are encountered various messages will be displayed telling you what discrepancies have been found.

- 6) The program checks if the data values provided are in accordance with the limitations defined for the type of analysis being run.
  - a. For weights it is checked that all weights are numeric values between 1 and 0. If non numeric values occur you will receive the message: "Weights contain blanks or non digit values. All values must be digits. If numbers outside the range of 1-0 occur you will receive the message: "Some weights are either larger than 1 or smaller than 0. All values must be between 1 and 0".
  - b. For CA it is checked that all values are positive integers, and that no objects nor variables have zero sums of values. CA cannot handle zero sums for neither rows nor columns (division by zero). If arrays with weights are not supplied you will be informed which objects and/or variables that have a zero sum, and you are asked to remove them. If arrays with weights have been supplied, CAPCA will simply remove those objects or variables that have zero sums. This difference in behaviour is due to CAPCA's use of the weighting arrays to carry out an automatic removal.
  - c. For PCA it is checked that all values are numeric and that no cells are left blank.
  - d. For MS it is initially checked if the variable types consist of "con" and "cat" exclusively. If this is the case then if continuous variables are indicated it is checked that all values are numeric and that no cells are left blank. If categorical variables are indicated it is checked that all values are either 1 or 0 (blanks being converted to 0).

When the program is halted you do not lose the information you have typed into the input form. The input form is a non-modal "floating form", meaning that you can make changes to the worksheet beneath without closing the form. You can thus correct errors as they are reported and quickly rerun the analysis without having to re-enter the settings.

Even when the analysis has completed the input form stays on with all its settings. You can thus inspect the result, find out if you need to change the number of Eigenvectors calculated, change weights, use transformations etc. and after having made the necessary changes quickly rerun the analysis.

When the analysis is running, two status boxes will appear at the bottom of the input form. The first is marked *Convergence measure* and the other *Number of iterations*. The convergence measure will start with a large value (negative or positive) that gradually will become smaller as the number of iterations grows. Calculation will stop when the convergence measure reaches a value of 0,5.

## **MATRIX OUTPUT**

The use of weights and the automated exclusion of objects and variables with sums less than required in a CA can alter the input data considerably. To allow inspection of the actual data sheet named *Matrix output* is created. It will contain a matrix of input data in the exact form they were supplied for analysis

The top row contains the names of the variables. The second row contains the classification of the variables. If no variable classification was supplied, this row will be empty. The first column of the sheet contains the names of the objects. The second column contains the classification of the objects. If no object classification was supplied, this column will be empty. Starting in cell C3 follows the block of values as analysed (with alterations by weights if supplied).

The first row beneath this block of values will be empty. The following row will show the weights for the variables. If no variable weights were applied, this row will be empty. The next row

will show the sum of variables. The final row will show the number of cells with non zero values for each variable.

The first column to the right of the block of values will be empty. The following column will show the weights for the objects. If no object weights were applied, this column will be empty. The next column will show the sum of objects. The final column will show the number of cells with non zero values for each object.

A second worksheet named *Seriation output* will be present if you in connection with a CA had checked the box *Seriate data set based on CA result*. This presents the data as analysed in a sorted order based on orthogonal projections of objects and variables onto second order polynomials through the objects and variables respectively on the two first principal axes. In terms of seriation this sorting is reliable only if the second order polynomials have a high degree of fit. Below the matrix the equations of the polynomials are given together with Pearson's squared R for goodness of fit. The value of this should preferably not be lower than 0.9.

Please note that if the worksheets *Matrix output* and *Seriation output* already exist they will be overwritten, and all data from the previous analysis will be lost. If you wish to keep the data for later use you should rename the worksheets.

## STATISTICS

When running an analysis, a new worksheet called *Statistics* is created. All statistics relating to an analysis, whether PCA, CA or MS, is written into this sheet. Please note that if the worksheet already exists it will be overwritten, and all data from the previous analysis will be lost. If you wish to keep the data for later use you should rename the worksheet.

### Statistics for PCA

The statistics worksheet contain information on normality of variables, covariance or correlation coefficients between variables, eigen values for calculated eigenvectors, variable loadings and object scores.

	Skewness	Kurtosis	Covariance matrix																					
Base width	-0,01	-0,34	2,35																					
Belly height	0,97	0,96	5,49	18,73																				
Belly width	0,70	0,54	2,25	8,41	5,41																			
Belly curvature	0,68	0,46	0,31	1,34	0,84	0,15																		
Shoulder height	0,02	-0,87	1,44	3,55	2,14	0,34	1,89																	
Shoulder width	0,36	-0,54	0,36	0,75	0,62	0,09	0,57	0,25																
Shoulder curvature	-0,13	0,16	0,00	-0,02	0,02	0,00	0,02	0,02	0,00															
Neck base height	1,95	4,18	0,34	0,38	1,08	0,12	0,31	0,21	0,03	1,22														
Neck base width	1,25	1,32	0,07	0,11	0,17	0,02	0,07	0,04	0,00	0,18	0,04													
Neck height	0,42	0,88	1,57	4,31	2,94	0,43	1,69	0,56	0,03	0,63	0,13	2,58												
Neck width	1,70	4,30	0,61	2,28	1,15	0,19	0,48	0,08	0,00	-0,07	0,00	0,73	0,48											
Neck curvature	1,89	4,65	0,04	0,17	0,11	0,02	0,04	0,01	0,00	0,02	0,00	0,08	0,03	0,01										

For each variable skewness and kurtosis is shown. Skewness is a measure that shows the degree of asymmetry around the mean value of a variable. It attains zero for perfect symmetry, has growing positive value with a growing asymmetric tail towards more positive values, and has growing negative value with a growing asymmetric tail towards more negative values. Kurtosis is a measure that shows whether the distribution of values is higher or lower than the normal distribution associated with the standard deviation. Positive values indicate a too high distribution and negative values a too low distribution.

Skewness is calculated as:

$$\frac{n}{(n-1)(n-2)} \sum_{j=1}^n \left( \frac{x_j - \bar{x}}{s} \right)^3$$

And kurtosis as:

$$\left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{j=1}^n \left( \frac{x_j - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

where s is the standard deviation of the variable and n its number of values. For both skewness and kurtosis the build in Excel functions are used.

If, as in the example above you have been running a PCA based on a covariance matrix you will find the covariance matrix to the right of the information on skewness and kurtosis. The covariance coefficients are calculated using the following formula:

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

If you are dealing with a covariance matrix the value of skewness and kurtosis really doesn't matter, but if you are dealing with a correlation matrix, they do because variables here should be reasonably normal distributed. If you have values of skewness and kurtosis between -1 and 1 you can consider the variables perfectly suitable for analysis. Only with values below -1 or above 1 you have reason to be critical, but still it may not be necessary to exclude variables. It depends on your overall assessment of their impact.

The information for a PCA based on a correlation coefficient looks much the same as the information associated with the covariance coefficient. The coefficient matrix, however, consist of correlation coefficients in stead of covariance coefficients, and the information on skewness and kurtosis values may have supplemental information on scale transformations.

	Skewness (ArcSin transformed)		Kurtosis (ArcSin transformed)		Correlation matrix																			
Base width	-03/(.27)		-08/(.67)		1,00																			
Belly height	.98/(.74)		1.01/(.16)		0,71	1,00																		
Belly width	1.11/(.51)		1.64/(.37)		0,44	0,83	1,00																	
Belly curvature	1.33/(-2.98)		2.52/( 14.17)		0,33	0,72	0,91	1,00																
Shoulder height	.51/(.62)		-.34/(.09)		0,54	0,47	0,55	0,60	1,00															
Shoulder width	.63/(-1.58)		.14/( 1.13)		0,36	0,31	0,48	0,45	0,81	1,00														
Shoulder curvature	.09/(.35)		.22/(-1.34)		-0,09	-0,08	0,14	0,08	0,14	0,51	1,00													
Neck base height	2.35/(-1.09)		8.33/( 1.03)		0,27	0,33	0,47	0,36	0,14	0,27	0,22	1,00												
Neck base width	2.85/(-.1)		13.88/(-1)		0,35	0,37	0,43	0,34	0,26	0,34	0,13	0,84	1,00											
Neck height	.36/(.7)		.7/( 1.25)		0,36	0,39	0,55	0,42	0,48	0,54	0,29	0,05	0,03	1,00										
Neck width	1.5/(-.63)		3.16/(-.09)		0,22	0,35	0,40	0,37	0,39	0,36	0,15	-0,15	-0,11	0,74	1,00									
Neck curvature	2.17/(.07)		6.09/(-1.55)		0,01	0,22	0,38	0,34	0,20	0,33	0,31	0,07	0,03	0,65	0,74	1,00								

If you have specified *No transformation*, skewness and kurtosis of the input data will be specified. If you have asked for a transformation - either *Log transformation* (ten based logarithm), *Ln Transformation* (natural logarithm) or *ArcSin transformation* (Arcus Sinus function) – you will still see skewness and kurtosis of the input data, but now followed in brackets by skewness and kurtosis of the transformed data.

When the analysis is based on correlation coefficients (Pearson's r) the matrix of these are shown to the right of the normality information.

The coefficients are calculated as:

$$r = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}}$$

Below the normality information there is a line telling you how many iterations it took before the algorithm converged to an acceptable level. Next follows the *Eigen values* for the calculated *Principal components* as well as the corresponding explanation percentages and cumulative explanation percentages.

Analysis completed after 50 iterations

	1. Principal component	2. Principal component	3. Principal component	4. Principal component	5. Principal component	6. Principal component	7. Principal component	8. Principal component	9. Principal component	10. Principal component
EigenValues	5,28	1,66	1,31	0,94	0,73	0,61	0,50	0,42	0,25	0,15
Explanation %	43,96	13,81	10,91	7,84	6,07	5,11	4,15	3,48	2,09	1,25
Cumulative Explanation %	43,96	57,77	68,69	76,53	82,60	87,72	91,86	95,35	97,44	98,69

Finally, the *Variable loadings* describing the original variables impact on the Principal components and the *Object scores* – the objects values on the principal components are shown.

Variable loadings

	1. Principal component	2. Principal component	3. Principal component	4. Principal component	5. Principal component	6. Principal component	7. Principal component	8. Principal component	9. Principal component	10. Principal component
Base width	0,32	-0,16	0,06	0,13	0,34	0,53	-0,42	0,04	0,20	-0,01
Belly height	-0,35	0,37	-0,14	-0,11	-0,02	-0,15	0,10	0,08	-0,27	-0,01
Belly width	-0,38	0,15	0,02	-0,09	0,27	0,36	0,13	0,05	-0,08	0,18
Belly curvature	0,36	-0,17	-0,08	0,07	-0,40	-0,36	0,03	-0,10	0,20	0,12

Object scores

	1. Principal component	2. Principal component	3. Principal component	4. Principal component	5. Principal component	6. Principal component	7. Principal component	8. Principal component	9. Principal component	10. Principal component
1	0,46	-0,23	-0,23	0,16	0,03	0,10	0,03	-0,19	-0,15	0,00
2	0,30	-0,13	-0,20	-0,01	0,02	0,01	0,08	-0,07	-0,03	0,00
3	-0,41	-0,13	0,12	-0,04	0,03	-0,12	-0,09	-0,17	-0,12	-0,01
4	-0,46	-0,19	0,16	-0,23	-0,08	0,06	0,05	0,03	-0,04	0,03

## Statistics for CA

The statistics worksheet contain information on inertia, eigen values for calculated eigenvectors, variable and object coordinates, information on quality of fit for variables, and information on quality of fit for objects.

The first line of information gives you the number of iterations it took before the algorithm converged to an acceptable level. This is followed by the total *Inertia* of the data (= sum of eigen values), the *Eigen values* for each of the calculated *Principal axes*, their percentage of the total inertia, and the corresponding cumulative percentages.

Analysis completed after 52 iterations  
Total inertia of data

	1. Principal axis	2. Principal axis	3. Principal axis	4. Principal axis	5. Principal axis	6. Principal axis	7. Principal axis	8. Principal axis	9. Principal axis	10. Principal axis
Eigen values	0,48	0,28	0,24	0,13	0,09	0,06	0,03	0,02	0,01	0,00
Explanation %	36,14	21,27	17,69	9,40	6,46	4,32	2,53	1,35	0,85	0,00
Cumulative Explanation %	36,14	57,41	75,10	84,50	90,96	95,28	97,81	99,15	100,00	100,00

The next two blocks of data gives the *Variable coordinates* and *Object coordinates* on the calculated principal axes. These coordinates are used for variable and object plots, and should primarily be viewed through their graphical representation, but they are listed as part of the statistics to allow user generation of plots or potentially other statistical handling of coordinates.

Variable coordinates

	1. Principal axis	2. Principal axis	3. Principal axis	4. Principal axis	5. Principal axis	6. Principal axis	7. Principal axis	8. Principal axis	9. Principal axis	10. Principal axis
Whipped cord	1,02	-0,60	-0,67	-4,65	-2,53	-1,17	0,39	0,14	-0,22	-0,66
Twisted cord	-0,57	0,86	2,64	-0,43	-0,01	0,10	-0,26	0,03	-0,20	0,36
Ordinary stab-and-drag	-1,90	-0,91	-0,18	0,02	0,19	-1,15	0,72	-2,26	2,19	0,93
Broad stab-and-drag	-2,49	-1,63	-0,79	-0,02	0,12	0,00	-1,23	-0,78	-3,77	2,20

Object coordinates

	1. Principal axis	2. Principal axis	3. Principal axis	4. Principal axis	5. Principal axis	6. Principal axis	7. Principal axis	8. Principal axis	9. Principal axis	10. Principal axis
Bønnerup	-1,80	-1,00	-0,62	0,05	0,06	0,01	-0,24	-0,12	-0,26	-1,36
Toftum A1	0,50	0,19	-0,45	-0,22	-0,80	0,60	0,23	-0,41	0,24	0,17
Toftum A31	0,97	-0,69	-0,06	-0,47	0,60	-0,36	0,15	0,49	-0,28	0,97
Stengade 2	0,24	1,27	-0,76	1,23	-1,33	-0,31	-0,75	-0,04	-0,28	-0,14

The remaining part of the statistics presents a number of useful statistics for the interpretation of the results, mainly based on the concepts of relative and absolute contribution to principal axes.

The two first blocks of diagnostic data are termed *Quality of fit in plots – variables* and *Quality of fit in plots – objects*. The first column in the blocks is marked *Mass %*, and shows the sum of counts for variables and objects respectively as percentages. This is in fact the column and row sum values that are used to calculate the table of expected values to which we compare the actual values. The following column marked *Inertia %* tells you how large a part of the total variation is accounted for by the individual variables and objects respectively as percentages. The remainder of these two blocks of diagnostic data consists of a number of columns (up to ten) showing the relative contributions of variables and objects to principal axes as plotted two by two (the sum of relative contributions for two axes plotted together). There are two conditions for data to appear in these columns. The first is of course that the principal axes to appear have been calculated. The second is that a sufficient number of axes have been calculated to allow the calculation of the relative contributions. If the latter is not the case you will find the message *Relative contributions could not be calculated* across the columns. If you meet this message and you want the information you should raise the number of principal axes to be calculated in the analysis.

Quality of fit in plots - variables

	Mass %	Inertia %	Fit on Principal axis 1 & 2	Fit on Principal axis 1 & 3	Fit on Principal axis 2 & 3	Fit on Principal axis 1 & 4	Fit on Principal axis 2 & 4	Fit on Principal axis 3 & 4	Fit on Principal axis 1 & 5	Fit on Principal axis 2 & 5	Fit on Principal axis 3 & 5	Fit on Principal axis 4 & 5
Whipped cord	3,09	9,37	4,39	4,64	2,53	71,03	68,92	69,17	23,28	21,17	21,42	87,81
Twisted cord	10,70	16,35	12,43	86,27	91,11	5,97	10,81	84,66	3,80	8,64	82,48	2,18
Ordinary stab-and-drag	5,80	9,69	25,98	21,31	5,04	21,13	4,86	0,19	21,33	5,07	0,39	0,21
Broad stab-and-drag	3,73	11,50	28,87	22,25	10,67	20,23	8,65	2,03	20,28	8,70	2,08	0,05

Quality of fit in plots - objects

	Mass*100	Inertia %	Fit on Pricipal axis 1 & 2	Fit on Pricipal axis 1 & 3	Fit on Pricipal axis 2 & 3	Fit on Pricipal axis 1 & 4	Fit on Pricipal axis 2 & 4	Fit on Pricipal axis 3 & 4	Fit on Pricipal axis 1 & 5	Fit on Pricipal axis 2 & 5	Fit on Pricipal axis 3 & 5	Fit on Pricipal axis 4 & 5
Bønnerup	16,18	23,54	64,13	54,94	20,98	49,09	15,13	5,94	49,10	15,14	5,95	0,09
Toftum A1	13,15	2,66	15,59	24,61	12,99	16,20	4,57	13,59	48,19	36,56	45,58	37,16
Toftum A31	8,67	4,27	41,39	27,67	13,96	34,08	20,37	6,65	38,01	24,30	10,58	16,99
Stengade 2	8,25	6,12	26,64	10,07	34,88	25,16	49,96	33,39	28,95	53,75	37,18	52,27

The next two blocks of diagnostic data shows the relative contributions of variables and objects respectively to the individual Principal axes. If not enough axes have been calculated to allow the calculation of the relative contributions the message *Relative contributions could not be calculated* is written across the columns.

*Relative contribution of a variable to a principal axis* is seen as the part of the inertia of the variable accounted for on a principal axis, where the sum of its contribution to all principal axes amount to 100%. It is computed as:

$$\frac{r_i f_{ik}^2}{\sum_{k'=1}^m r_i f_{ik'}^2}$$

where  $r_i$  is the row sum for the  $i^{\text{th}}$  object,  $f_{ik}$  is the coordinate value for the  $i^{\text{th}}$  object on the  $k^{\text{th}}$  principal axis,  $m$  is the number of principal axes, and  $k'$  is a principal axis.

Relative contributions of variables to principal axes

	1. Principal axis	2. Principal axis	3. Principal axis	4. Principal axis	5. Principal axis	6. Principal axis	7. Principal axis	8. Principal axis	9. Principal axis	10. Principal axis
Whipped cord	3,25	1,14	1,39	67,78	20,03	4,32	0,48	0,06	0,15	1,39
Twisted cord	3,80	8,64	82,48	2,18	0,00	0,12	0,78	0,01	0,49	1,51
Ordinary stab-and-drag	21,13	4,86	0,18	0,00	0,21	7,74	2,99	29,75	28,07	5,08
Broad stab-and-drag	20,23	8,65	2,03	0,00	0,05	0,00	4,94	1,97	46,33	15,81

*Relative contribution of an object to a principal axis* is seen as the part of the inertia of the object accounted for on a principal axes, where the sum of its contribution to all principal axes amount to 100%. It is computed as:

$$\frac{c_j g_{jk}^2}{\sum_{k'=1}^m c_j g_{jk'}^2}$$

where  $c_j$  is the column sum for the  $j^{\text{th}}$  variable,  $g_{jk}$  is the coordinate value for the  $j^{\text{th}}$  variable on the  $k^{\text{th}}$  principal axis,  $m$  is the number of principal axes, and  $k'$  is a principal axis.

Relative contributions of objects to principal axes

	1. Principal axis	2. Principal axis	3. Principal axis	4. Principal axis	5. Principal axis	6. Principal axis	7. Principal axis	8. Principal axis	9. Principal axis	10. Principal axis
Bønnerup	49,05	15,09	5,90	0,04	0,05	0,00	0,86	0,20	0,98	27,83
Toftum A1	13,61	1,98	11,00	2,58	34,58	19,57	2,92	9,00	3,15	1,60
Toftum A31	27,55	13,84	0,12	6,53	10,46	3,92	0,68	7,06	2,34	27,50
Stengade 2	0,92	25,72	9,15	24,24	28,03	1,52	8,86	0,02	1,24	0,30

The final two blocks of diagnostic data shows the absolute contributions of variables and objects respectively to the individual principal axes.

*Absolute contribution of a variable to a principal axis* is seen as the part of the inertia of the principal axes accounted for by the variable where the sum of the contribution of all variables to the principal axes amount to 100%. It is computed as:

$$\frac{r_i f_{ik}^2}{\sum_{i'=1}^n r_{i'} f_{i'k}^2}$$

where  $r_i$  is the row sum for the  $i^{\text{th}}$  object,  $f_{ik}$  is the coordinate value for the  $i^{\text{th}}$  object on the  $k^{\text{th}}$  principal axis,  $n$  is the number of objects, and  $i'$  is an object.

Absolute contributions of variables to principal axes

	1. Principal axis	2. Principal axis	3. Principal axis	4. Principal axis	5. Principal axis	6. Principal axis	7. Principal axis	8. Principal axis	9. Principal axis	10. Principal axis
Whipped cord	2,37	0,71	0,86	63,41	23,81	6,51	0,65	0,08	0,14	0,79
Twisted cord	4,46	8,58	82,01	3,28	0,00	0,30	1,68	0,02	0,72	1,39
Ordinary stab-and-drag	16,10	3,13	0,12	0,00	0,26	12,17	4,17	37,68	26,89	3,03
Broad stab-and-drag	21,08	7,62	1,79	0,00	0,09	0,00	9,42	3,42	60,71	12,91

*Absolute contribution of an object to a principal axis* is seen as the part of the inertia of the principal axes accounted for by the object where the sum of the contribution of all objects to the principal axes amount to 100%. It is computed as:

$$\frac{c_j g_{jk}^2}{\sum_{j'=1}^n c_{j'} g_{j'k}^2}$$

where  $c_j$  is the column sum for the  $j^{\text{th}}$  variable,  $g_{jk}$  is the coordinate value for the  $j^{\text{th}}$  variable on the  $k^{\text{th}}$  principal axis,  $n$  is the number of variables, and  $j'$  is a variable.

Absolute contributions of objects to principal axes

	1. Principal axis	2. Principal axis	3. Principal axis	4. Principal axis	5. Principal axis	6. Principal axis	7. Principal axis	8. Principal axis	9. Principal axis	10. Principal axis
Bønnerup	77,65	33,74	28,26	0,21	0,34	0,04	8,44	2,26	16,85	47,25
Toftum A1	0,56	0,11	1,37	0,35	5,82	5,98	0,74	2,62	1,40	0,07
Toftum A31	2,18	1,55	0,03	1,72	3,39	2,31	0,33	3,95	2,00	2,33
Stengade 2	0,18	7,25	5,53	16,05	22,91	2,25	10,91	0,03	2,68	0,06

NB! As relative contributions are calculated as percentages of representation on all principal axes they cannot be computed precisely unless all principal axes have been computed. As these in a CA often are many it may take a very long time for the analysis to complete, and further if there is no variation left on the last principal axes the calculation algorithm may misbehave. As we are interested in the first few principal axes only and as it is highly unlikely that the relative contribution of variables or objects to the first principal axes will be significantly affected by what happens on the last principal axes, we may without major problems calculate the relative contributions based on 90% of the total variation. This is set as the cut off value for calculating relative contributions.

## Statistics for MS

The only statistics you will receive from a metric scaling is the coordinate for the objects.

### GRAPHICAL REPRESENTATION

For all three types of analysis the best way to view the result is through a plot of objects and variables against the principal axes/components. To create such diagrams you open an additional form, where you can set the parameters for the plots. This is done by pressing the button on the input form marked *Go to diagrams*. There are a couple of things you should be aware of here.

Although the coordinates used to plot objects and variables in relation to the principal axes/components has been written in the *Statistics* worksheet, CAPCA needs some additional information stored in memory to create the plots. This means that plots can only be created for a newly run analysis. If you have closed down Excel, and reopened it you won't be able to create the plots directly from the *Statistics* worksheet, and you will find that the button *Go to diagrams* on the input form is inactive until you have run an analysis.

When you open the diagrams form, the input form disappears. It is not closed, however, only hidden. If you press the *Back to calculations* button on the diagrams form the input form will reappear with all the data ranges and settings intact. You can then rerun the analysis with what ever changes in data or settings needed.

### Settings for PCA

For a PCA the diagrams form will appear as follows:

The screenshot shows a dialog box titled "CAPCA version 2.0 - © Torsten Madsen 2005-2007". The main heading is "Select plots". There are two columns of checkboxes. The left column has a header "Axis 1" and five rows labeled "Axis 2", "Axis 3", "Axis 4", and "Axis 5". Each row has two checkboxes, one under "Axis 1" and one under the corresponding "Axis" label. The right column has four checkboxes labeled "Show variable names", "Show object names", "Show object classification", and "Show variable classification". At the bottom, there are two buttons: "Make diagrams" and "Back to calculations".

You can select plots involving up till the first five principal axes in any combination you wish by placing checkmarks in the appropriate boxes to combine axes. If you have calculated less than five axes only the number of axes calculated will be available. For each combination of axes you can have two types of plots: *Variable plots* and *Object plots*.

You can select if variable and/or objects names should appear in the plots. If object and/or variable classifications is supplied, classifications can be shown in the plots with different signatures. If classifications have not been supplied the check boxes will be disabled.



## Settings for CA

For a CA the diagrams form will appear as follows:

You can select plots involving up till the first five principal axes in any combination you wish by placing checkmarks in the appropriate boxes combining axes. If you have calculated less than five axes only the number of axes calculated will be available. For each combination of axes you can have three types of plots: *Variable plots*, *Object plots* and *Combined plots*. The latter is a plot where variables and objects both occur in the same plot.

You can select if variable and/or objects names should appear in the plots. If object and/or variable classifications is supplied, classifications can be shown in the plots with different signatures. If

classifications have not been supplied the check boxes will be disabled. You can also have objects and variables shown with different signatures based on a class division of their Inertia. For technical reasons you cannot use this facility together with the classifications. So if you select *Show object Inertia*, *Show object classification* will automatically become unselected, and vice versa.

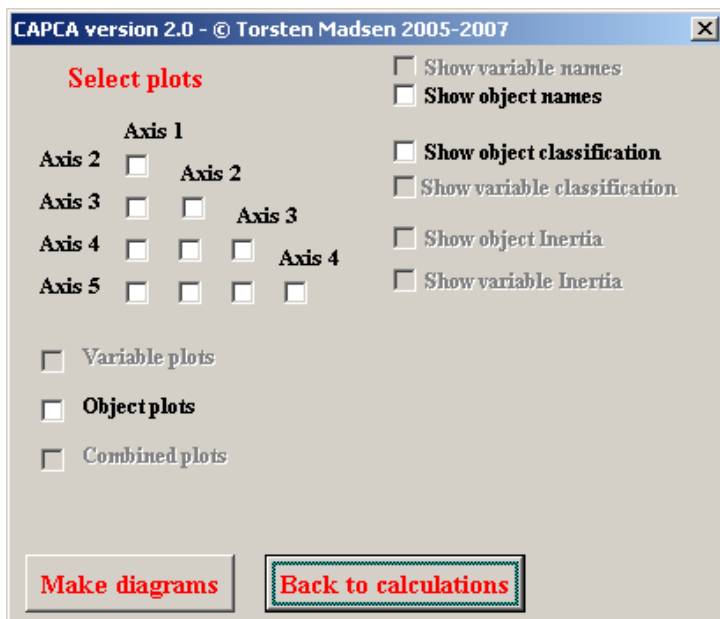
If relative contributions have been calculated, two checkboxes allowing you to remove objects and variables with low relative contributions from the plots will be enabled. Otherwise they are disabled.

If enabled, and you checkmark the boxes, two input boxes opens in which you must write the values below which the objects and variables will not be shown. In order to understand the function of this facility you should consider the following. The principal axes will represent the variables and objects differently. An object may have its variation badly represented on the first and second axes, but very well represented on the third axes. However bad its representation on the first and second axis is,

it will still be there with coordinates, and it will appear in the plots. Its position in the plots may be rather hap hazardous (though mostly close to the centre) in relation to what the axes show and it may indeed be disturbing to the interpretation of the axes. The relative contribution is the statistics that tells us how well an object or a variable is represented on a principal axes, and hence we may use it to eliminate objects and variables with a bad representation. Note, they are eliminated from the plots – not from analysis.

## Settings for MS

For a MS the diagrams form will appear as follows:



There are few available options. Object plots are thus the only type of plot possible. On these plots you can have the object names shown, and if you have provided an object classification, the objects in the plot can be grouped by this.

### EDITING AND SAVING THE PLOTS

The plots are created entirely using the Microsoft Excel Chart facility. This results in both advantages and disadvantages. To take the main disadvantage first, any Excel chart is linked to data in a worksheet. Thus as the first step to create the plots, CAPCA fills up a worksheet called *Chartdata* with the information needed to produce the plots. The plots are depending on *Chartdata*, so if this is deleted the content of the plots will disappear. A deletion of *Chartdata* will happen the next time you press the *Make diagrams* button, and actually the same is true with the plots, as these also have standard names. Thus if you want to keep the plots you either have to rename both the plots and *Chartdata*, or better move them to a new workbook. It is important to remember here that you cannot produce plots piecemeal and successively move them to another workbook. The worksheet *Chartdata* is unique for each run of *Make diagrams*, and it must follow the plots.

The advantage of using the Excel Chart facility is that the charts are fully editable once created. You can use all the standard editing facilities associated with Excel Charts including moving of individual labels to make plots with variable and object names more readable. It is certainly worth while to spend some time to familiarize yourself with the editing potentials.

If you wish to save a plot independently of data in a worksheet you can do this by copying a plot to the clipboard and from there move it to other programs. One such type of programs is a bitmap editor like Adobe Photoshop. Saved from here plot ends up in a bitmap format like jpeg or tiff. The advantage is that the plots are stored in an easily accessible format, shared by all. One drawback is that the resolution will be fixed and screen depended, which from a publication point of view is bad. Another is that once converted to bitmap it is no longer possible to edit the plots.

Another option is to copy the clipboard content to a vector editor like Adobe Illustrator. This is the professional solution if publication is the aim. All the elements of the plot will be fully editable,

and the editing facilities go far beyond what you can do within Microsoft Excel. The drawback is that very few has access to a program like Adobe Illustrator, and you have to have a fair amount of knowledge about the program to use it properly.

### **Literature**

Gower, J.C. 1971 A general coefficient of similarity and some of its properties. *Biometrics* 27, p. 857-74.

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.