Multivariate analyses in Excel using PCA, CA and MS
CAPCA Version 3.1

Torsten Madsen ©2016

Introductory comments

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) fully integrated within Excel. The program reads the data from your worksheets and produces additional worksheets with tables and statistics showing the result of the analysis as well as a number of charts to illustrate it.

The background for CAPCA was a DOS based program in PACAL performing PCA and CA analyses that I wrote in 1989 called Kvark. The program was abandoned as Erwin Scollar's successful program Basp and later WinBasp developed during the first half of the nineties. Both Kvark and Basp/Winbasp, however, were hampered by tedious data input procedures and a lack of efficient import facilities for data.

The advantage of a program for multivariate analyses within Excel appeared rather obvious, and around 2002 I made a first attempt to create CAPCA. The program was primarily used internally at the department, although it was disseminated to a few colleagues as well. In 2006, when I withdrew from the university, I decided to make it generally available. It was launched as version 1, followed in 2007 by version 2 (version 2.1 in 2010 and version 2.3 in 2012). Version 3.0 from 2014 presented a major renewal and improvement of the user interface. The present version 3.1 features some needed improvements to the way you can enter data for metric scaling, as well as some improvements to the way you can present and view the results of the analyses.

CAPCA is written using the built in VBA programming language in Excel. The version of Excel used for version 3.1 is 14 (Office 2010). I have not had the possibility to test the program using newer versions, but it ought to work. I would appreciate if you inform me, should you encounter difficulties running the program, if you come across regular bugs, or indeed if you have any suggestions of improvement. The code is not password protected, and you are welcome to browse through it and make changes if you like. Should you make regular improvements to the program, I would appreciate if you inform me.

Installing CAPCA

Installing CAPCA on your computer is simple. The only precondition is that you have Microsoft Excel version 12 (Office 2007) or later. The program is distributed as an Excel add-in program file (.xlam file) that is compatible with Office 2007 or later.

CAPCA Version 3.1.xlam

It is customary to place add-in files in 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*). The directory may be hidden. In that case, you need to change the settings in the file browser allowing you to see hidden directories. It is not compulsory, however, to place the file in this directory. You may in fact place it wherever you want, but Excel will always look first for an add-inn file in the above mentioned directory.

The procedure for installing in Office 2010 is the following (it differs slightly from office 2007):

If you already have CAPCA installed you should remove the add-inn file (e.g. CAPCAversion3.03.xlam) before you open Excel. When you open Excel, it will inform you that the file is missing. Next, you activate the "Files" page in Excel and select settings. In the menu that appear you select Add-Ins, and in the subsequent Add-Ins dialog box you select "Manage Excel Add-ins" at the bottom of the form and press Go. You will see the name of the removed file checked in the add-inn list. You un-check it and let Excel remove the reference. Finally, you press browse to find "CAPCA Version 3.1.xlam" and activate it.

If you do not have a previous version of CAPCA installed, you go directly to the "Files" page in Excel and select settings. In the menu that appear you select Add-Ins, and in the subsequent Add-Ins dialog box you select "Manage Excel Add-ins" at the bottom of the form and press Go. You then press browse to find "CAPCA Version 3.1.xlam" and activate it.

Potential problems installing CAPCA.

Hopefully you will not be met with problems when trying to run CAPCA, but with new versions of Microsoft Windows and Microsoft Office they may appear. If you experience that a newly installed program crash, please report it.

Previously, there were problems with outdated versions of a Microsoft library file named REFEDIT.DLL. If you are using Office 2007 you will find the file in the library C:/Program files (x86)/Microsoft Office/Office12. It should have a date of 26.02.2009. The original version of the file for Office 2007, dated 26.10.2006, resulted in fatal run time errors.

The version of the file used here is dated 13.03.2010. It is placed in C:/Program files (x86)/Microsoft Office/Office14.

Preparing data for CAPCA

All data to be used in the analyses must be placed in an Excel worksheet. Seven sets or elements of data are recognised: *Values*, *Object names*, *Variable names*, *Object classes*, *Variable classes*, *Object weights* and *Variable weights*.

Values

The set of values consist of a rectangular table of numbers, blanks or strings. In computing terms it is a two-dimensional array (*Values(i,j)*), where *i* is the number of entries in the first dimension and *j* is the number of entries in the second dimension of the array. In Excel it equals a rectangular block of cells. Thus, if *i* = 28 and *j* = 18 it could be a block of cells with the reference *C3:T30*.

The type of values that you can enter into the table depends on the type of analysis. The rules differ for a PCA, a CA and a MS analysis. In connection with the chapters, describing each of these analyses you will find detailed information on what kind of values you can use. There is all reason to study this information carefully, because most failures to make the program work relates to illegal values in the data set.

Object names and Variable names

Object names and Variable names consist of a row or a column of text strings. In computing terms they are represented by one-dimensional array (*ObjectNames*(*i*) or *VariableNames*(*j*)), where *i* and *j* are the number of entries in the arrays. In Excel it equals a horizontal or vertical stripe of cells. Thus, if *i* = 28 it could be represented by a series of cells with the reference B3:B30, and if j = 18 it could be represented by a series of cells with the reference C2:T2.

Each name can consist of letters or digits alone or in any combination, and there is no limitation to the length of the names. Missing names (blank cells) are not allowed, and each name must be unique within the set of names. Both Object names and Variable names are mandatory data.

Object classes and Variable classes

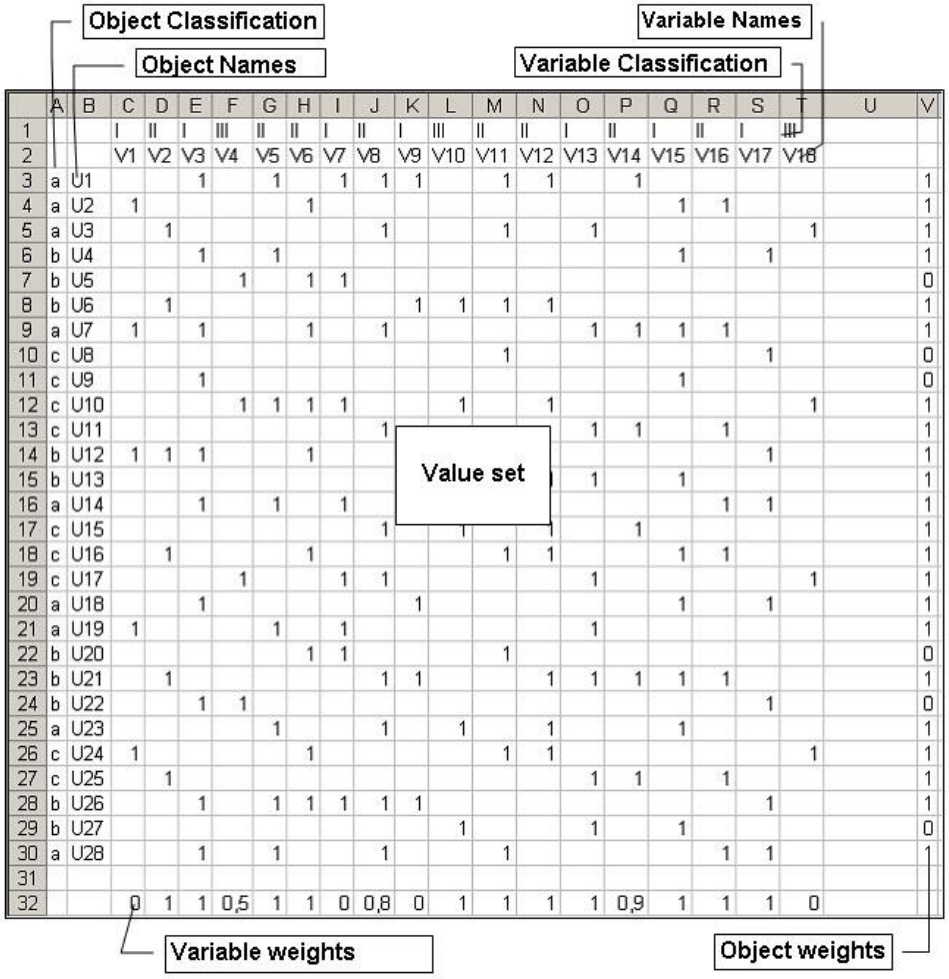
Object classes and Variable classes consist of a row or a column of text strings. In computing terms they are represented by one-dimensional array (*ObjectClasses*(*i*) or *VariableClasses*(*j*)), where *i* and *j* are the number of entries in the arrays. In Excel it equals a horizontal or vertical stripe of cells. Thus, if i = 28 it could be represented by a series of cells with the reference A3:A30 and if j = 18 it could be represented by a series of cells with the reference C1:T1

Each class name can consist of letters or digits alone or in any combination, and there is no limitation to the length of the class names. Missing names (blank cells) are not allowed. There is no limit to the number of different class names in the sets of object or variable classes. All class names could be unique or they could be identical, but of course it is only meaningful with a number of different class names somewhere in between. Object and variable classes are optional data.

Object weights and Variable weights

Object weights and Variable weights consist of a row or column of numbers. In computing terms it is a one-dimensional array (*ObjectWeights*(*i*) or *VariableWeights*(*j*)), where *i* and *j* are the number of entries in the arrays. In Excel it equals a horizontal or vertical stripe of cells. Thus, if *i* = 28 it could be represented by a series of cells with the reference V3:V30 and if *j* = 18 it could be represented by a series of cells with the reference C32:T32.

Each weight must be either 0 or 1, but for some analyses it can also be a digit between these two values. The weights are used differently in PCA CA and MS. For information about how they may be used you are referred to the description in connection with the individual analyses. Object weights and Variable weights are optional.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1			I	II	I	III	II	II	I	II	I	III	II	II	I	II	I	II	I	III		
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
15	b	U13													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
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	0		

Figure 1. For good practice, use a standardised layout of data for entry into CAPCA.

Organising the worksheet

Although you can place the elements of data wherever you want in a worksheet or indeed in different worksheets if you wish, it is customary and good practice with an organised setup as the one shown in Fig. 1. Here, object names and object classes to the left and variable names and variable classes above flank the values. Further, object weights are placed to the right of the values and variable weights below them.

It is not mandatory to have objects in the rows and variables in the columns of the value set, but it is certainly good practice. It is imperative, however, that the program knows the orientation of the value set, so if you have objects in the columns and variables in the rows you must remember to change the default setting in the data entry page of the main form.

Entering data into CAPCA

CAPCA has one main form and one auxiliary popup form. The main form has five pages: a data entry page, a page for running CA, a page for running PCA, a page for running MS, and a page for displaying graphics. Each page, I discussed in separate chapters.

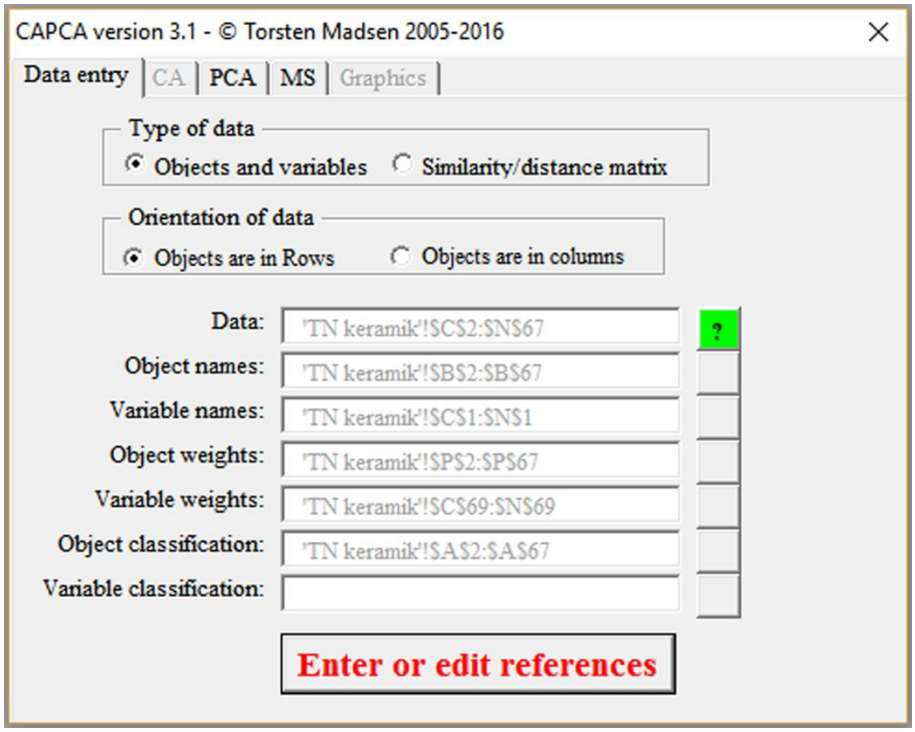


Figure 2. CAPCA main form, data entry page.

The first time you open CAPCA in a workbook, the reference controls will be blank, but as soon as you have entered references, these will be remembered, and whenever you open CAPCA again in that particular workbook the reference controls are filled in with the last used references (Fig. 2).

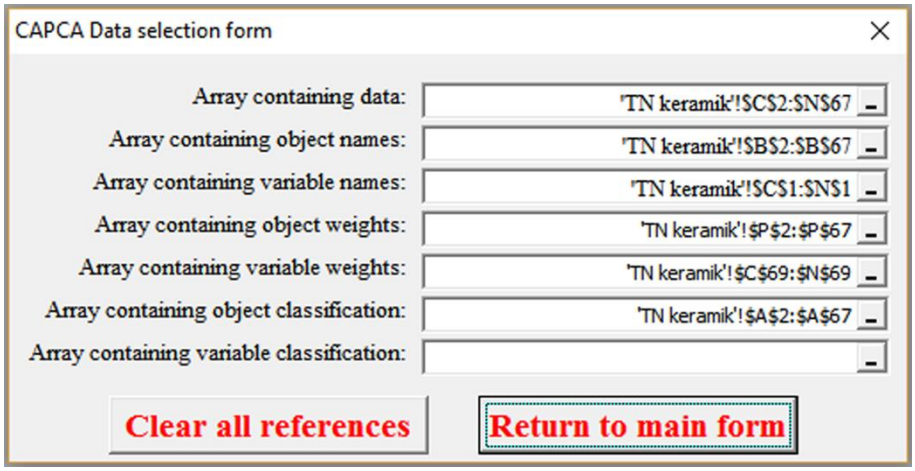


Figure 3. CAPCA data selection form.

Recording of references does not take place in the main form, but through an auxiliary popup form opened by pressing the button “Enter or edit references” (Fig. 3).

The popup form has seven controls through which you can enter references for the seven types of data categories. To enter a reference you either place the cursor in the reference field (mark the existing reference if there is one), and then mark up the relevant area of data in the sheet, or you click on the small square box to the left of reference. Doing this will make the popup form disappear leaving the one control in use visible only (Fig. 4). This makes it easier to see the data in the sheet and mark them. When done you click the small rectangular box again, and the form will reappear.

If you choose to fill in references in the controls manually, you must remember to enter the sheet name as part of the reference. In contrast to previous versions of CAPCA, references without a sheet name are not allowed.

Figure 4. RefEdit control during reference recording.

When you press “Return to main form” the references are transferred to the main form and the popup form disappears.

You may well wonder why the reference recording is done through an auxiliary form and not directly from the main form. The reason is that the RefEdit control does not work when embedded in a multipage form. It summarily and ungracefully crashes the program when activated. It is not mentioned in the documentation of the control, but a few references to this behaviour can be found on the web.

Option controls on the data entry page

There are two option controls on the data entry page. The first sets the *type of data*, the second sets the *orientation of data*.

Type of data here refers to data as a table of objects described through variables versus data as a table of coefficients describing the interrelationship between objects. The first is the default, and probably the only you will ever user. Similarity coefficients are used for MS only, and even here most users will probably choose to enter data as objects and variables and let CAPCA worry about the calculation of similarity coefficients.

Orientation of data refers to whether *objects are in rows* (and variables in columns) or *objects are in columns* (and variables in rows). The first is the default, and you are urged to keep this convention. If your objects are in columns the data set will be transposed before analysis, and eventually you will find that in all tabular output of data the objects have ended up in the rows. If you have chosen similarity coefficients as *type of data* the *orientation of data* option will be disabled as it is irrelevant.

Validation of data

A validation process is carried out continuously, triggered by almost any action in the main form. The aim of the process is to ensure that the referenced data are suitable for analyses and to determine what limitations there might be on these analyses. The result of the validation process is fed back to the main form, where you can access it through the information buttons that are placed to the right of the reference fields.

The information buttons has three states of appearance: neutral, green or red. A neutral button has the same colour as the background of the form, and nothing will happen if you press it. A green button or a red button will display a message box if you press it. The message box will display information related to the data referenced in the adjoining reference field.

A green button indicates that data with some restrictions can be analysed. Thus, in fig. 2 you can see that despite a green button and no red buttons the page for performing CA has been disabled. Pressing the green button, you would find out why CA is not available.

A red button indicates that something in the data referenced hinders any analysis. In Fig. 5 there are two red buttons each triggered independently by invalid data blocking an analysis. Pressing the buttons you will find information that explains what the problems are.

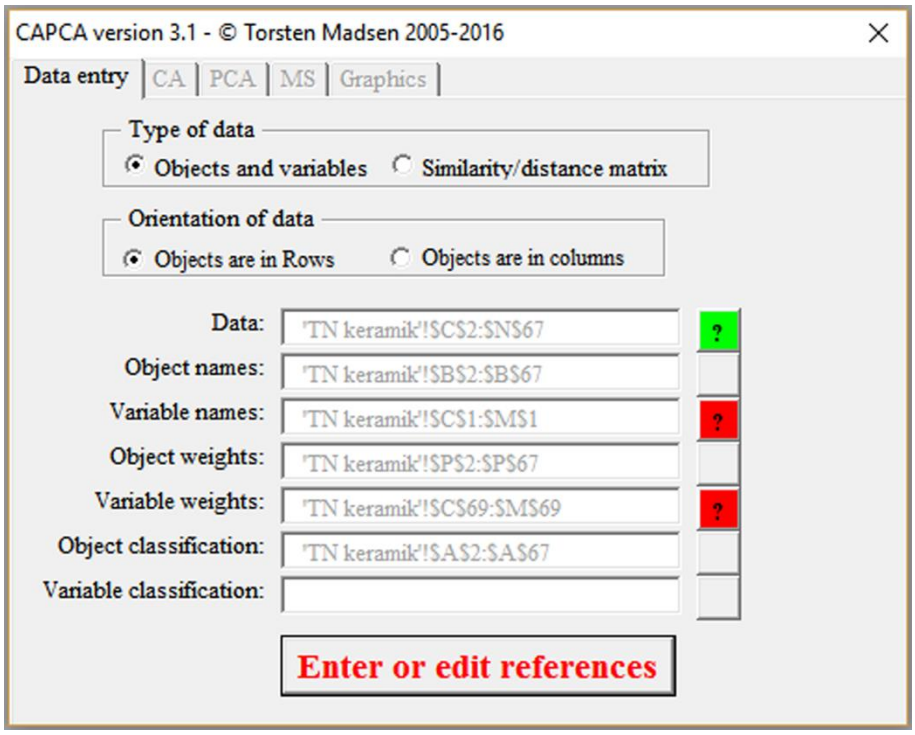


Figure 5. CAPCA main form, data entry page. Red info buttons signals invalid data. All pages other than the data entry page has been disabled.

The following list gives an overview of the messages you can encounter in connection with the various data sets:

Value set

- You must supply a complete reference, including the sheet name - eg. Sheet!cells.
- The data set contain 38 objects and 18 Variables.
- The data set contain blanks. Blanks can be used in CA only. The blanks will automatically be substituted by zeroes.
- The data set contain values that are not integers. The data cannot be used for CA analysis.
- The data set contain values that are negative. The data cannot be used for CA analysis.
- Log transformations in PCA are not available for variables with negative data. You may use ArcSin transformation in stead.
- The data set contains sums across object values that are zero. If you use a CA analysis these objects will be eliminated from the analysis.

- The data set contains sums across variable values that are zero. If you use a CA analysis these variables will be eliminated from the analysis.
- The data set contains non numeric data. It cannot be used for PCA or CA.

in MS

A separate set of messages apply to similarity coefficient matrices

- The data set contains question marks indicating missing values useable only for MS.
- The object weights supplied are ignored as data consist of similarity coefficients.
- The variable weights supplied are ignored as data consist of similarity coefficients.
- The variable classes supplied are ignored as data consist of similarity coefficients.
- The data set contains variables with a mix of numeric and non-numeric data. It is not allowed in MS.
- The similarity matrix contains non-numeric data. It cannot be analysed.
- Data appears to constitute a similarity or distance matrix. Please check the appropriate option.
- It is not a symmetric matrix. It cannot be used as a similarity coefficient matrix.
- Data contain negative data. Only positive values are allowed in a similarity coefficient matrix.
- The diagonal values are not identical. They must be in a similarity coefficient matrix.
- Off-diagonal values are larger than the diagonal values. They cannot be in a similarity coefficient matrix.
- Values are not mirrored around the diagonal. They must be in a similarity coefficient matrix.

Object names and Variable names

- You must supply a complete reference, including the sheet name - eg. Sheet!cells.
- The reference to object names must be one-dimensional. Names must be in one row or one column exclusively.
- The reference to variable names must be one-dimensional. Names must be in one row or one column exclusively.
- There are ## objects in your dataset and ## object names in the names array. The number must be the same.
- There are ## variable in your dataset and ## variable names in the names array. The number must be the same.
- One or more names are blank. A name must contain letters and or numbers.
- There are duplicate names. All names must be unique. The following are duplicates:
- You must provide names for the objects to run an analysis.
- You must provide names for the variables to run an analysis.

Object classes and Variable classes

- You must supply a complete reference, including the sheet name - eg. Sheet!cells.
- You must provide names for the objects to run an analysis.
- You must provide names for the variables to run an analysis.
- The reference to object classes must be one-dimensional. Classes must be in one row or one column exclusively.
- The reference to variable classes must be one-dimensional. Classes must be in one row or one column exclusively.
- There are ## objects in your dataset and ## object classes in the class array. The number must be the same.
- There are ## variables in your dataset and ## variable classes in the class array. The number must be the same.
- One or more class names are blank. A class name must contain letters and or numbers.

Object Weights and Variable weights

- You must supply a complete reference, including the sheet name - eg. Sheet!cells.

- The reference to object weights must be one-dimensional. Weights must be in one row or one column exclusively.
- The reference to variablet weights must be one-dimensional. Weights must be in one row or one column exclusively.
- There are ## objects in your dataset and ## weights in the weights array. The number must be the same.
- There are ## variables in your dataset and ## weights in the weights array. The number must be the same.
- Weights are found that contain blanks or non digit values. All values must be digits.
- Some weights are either larger than 1 or smaller than 0. All values must be between 1 and 0.
- In connection with MS only weights of 1 or 0 are valid. All other values will be converted to 1.

Calculating eigenvectors – number and precision

All three types of analyses in CAPCA are based on finding the eigenvectors (principal components or principal axes) of a matrix of values. You provide the values – the data, but depending on which type of analysis you select – CA, PCA or MS, certain changes are made to data prior to analysis. The actual calculations – the Eigen-decomposition, or in a broader sense the single value decomposition – are, however, the same for all three types of analyses. The calculation of Eigenvectors in CAPCA is based on an iterative algorithm published by Wright (1985).

You can calculate as many eigenvectors as the smallest dimension of your data. If there are fewer variables than objects the maximum number equals the number of variables. If there are fewer objects than variables the maximum number equals the number of objects. The default number of eigenvectors calculated is 3, but you can set any number for each of the three types of analysis. If you enter a number higher than the highest possible it is reduced to that number. If you enter a number smaller than the lowest possible (2) it is altered to the default number 3. The same applies if what you write cannot be interpreted as a number. Changing the number of eigenvectors to be calculated is done in the main control form on the individual pages for CA, PCA and MS.

It is not possible, at least not for larger matrices, to calculate the eigenvectors directly, simply because there are too many unknown variables in the equation. Therefore, general algorithms to find eigenvectors and corresponding eigenvalues are iterative. An iterative method for calculating eigenvectors and eigenvalues typically works through a sequence of guesses on appropriate eigenvectors and eigenvalues. After each guess it is checked how close the suggested values come to solve the equation. The result is used to correct the next guess in a direction that will give an even better fit. To control the sequence of guesses a convergence value is calculated that continuously monitor how close to a perfect fit the guesses are. When the convergence value has become sufficiently small and has reached a preset stop value the iterative process terminates.

The iterative method is not meant to, and indeed cannot find *the true* eigenvectors and eigenvalues. The result is always an approximation, and if we set our stop value too small we may occasionally experience that the iteration cannot find a satisfactory approximation. Therefore, it is necessary to have an alternative way to stop the iteration. In CAPCA this is done by stopping the iterative process if it exceeds 1000 loops.

It is important to note that the growing stability in the calculation of eigenvectors and eigenvalues expressed through the falling convergence value does not happen equally across all calculated eigenvectors. In general, stability is reached progressively through the sequence of eigenvectors/principal axes, and indeed it is questionable if stability can ever be reached for the last eigenvectors in a large value set.

This implies that if you are interested in the first two-three eigenvectors/principal axes only, you need not have a very small stop value. If, on the other hand, you want to have all eigenvectors calculated you need to set a small stop value. In CAPCA you control the precision through an option box with the choice of low, medium or high precision, where medium is the default setting. An example can illustrate the influence of these three settings.

93 objects with 41 were analysed using a CA with 20 eigenvectors/principal axes required. With low precision the analysis took 264 iterations to converge. With medium precision the number was 320 and with high precision it was 451. If we compare the eigenvectors and eigenvalues calculated with the three precisions we find that with low precision compared to high precision there is no deviation for the first 8 eigenvectors, while for the last 12 there is an irregular pattern of deviation up to 0.000005. If we compare medium precision to high precision there is no deviation for the first 12 eigenvectors, while for the last 8 there is an irregular pattern of deviation up to 0.000001.

Running a PCA

Fig. 6 shows the PCA page of the main control form. This page is enabled if the validation of data has shown that a PCA can be run with the data provided. Otherwise it will be disabled. The page provides an option box for setting the input format – whether data should be analysed through correlations coefficients or covariance coefficients (se below). Further, the page contain controls for setting the number of eigenvectors to be calculated and the precision with which they should be calculated (se the chapter on *Calculating eigenvectors – number and precision*). The following two check boxes controls if weights for objects and variables should be applied. They are only enabled if you have provided weights, and you must actively check them to apply the weights you have supplied. The following four info-fields inform you on how many objects and variables are being analysed and how many (if any) objects and variables have been excluded by user weights (see below). If exclusions occur a green button next to the number becomes available. Pressing this will open a message box giving the names of the objects and/or variables excluded. Finally, when the analysis has been completed, the number of iterations will be displayed.

Acceptable values for PCA analysis

All values in the provided data set must be numbers. They can be either integers or real, negative or positive, and include zero. Text is not allowed, and the only special characters allowed are (-) (.) and (,). They are interpreted according to the national setup of your computer (e.g. -2.455,23 in a Danish setup compared -2,455.23 in an English setup). Remember that Excel as a rule will align everything it reads as a number to the right and everything else to the left. Anything you see aligned to the left in your data in the worksheet will trigger an error.

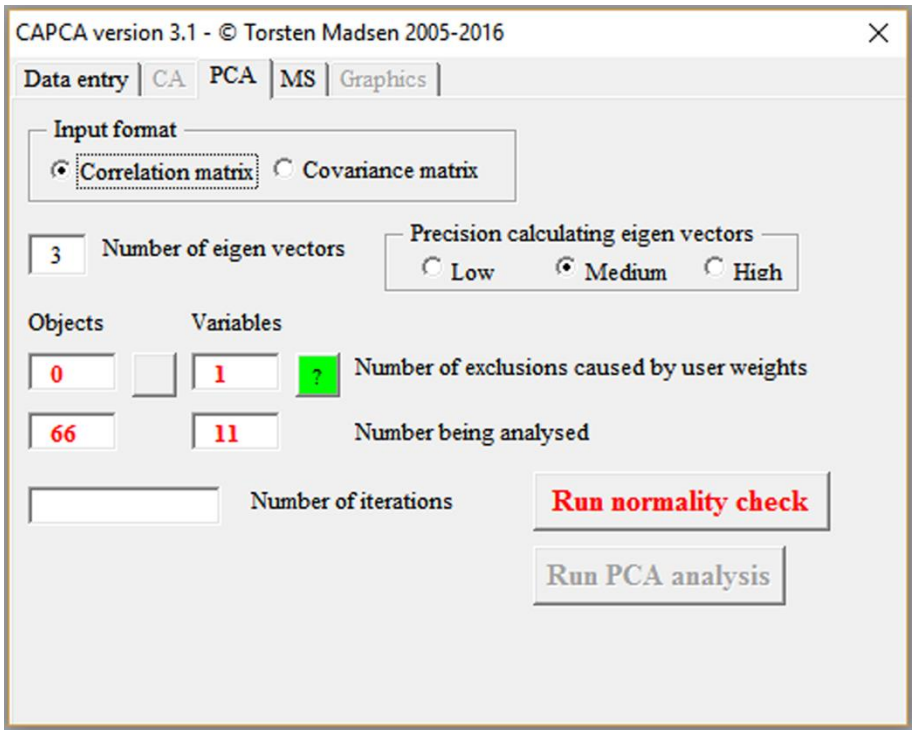


Figure 6. The PCA page of the CAPCA control form

Setting the input format

You can choose whether the PCA should be based on a correlation matrix (default) or a covariance matrix. In both cases alterations are made to the data you provide, but in different ways and with different results.

With the covariance matrix the values of a variable are altered by subtracting the mean value of the variable and divide by the square root of the total number of values minus one.

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

In this way the values of a variable will be centred on its mean value, but relatively speaking, size differences between variables are not affected. If you have a variable of length and a variable of thickness, the sum of the values of the first will still be larger than the sum of values of the latter.

With a correlation matrix the values of a variable are altered by subtracting 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.

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

In this way the values of a variable will not only be centred on its mean value, but the magnitude of the distribution around the centre will be altered (standardised) to unity as well. Thus all variables will become of the same size. If you have a variable of length and a variable of thickness, the sum of the values of the first will be the same as the sum of values of the latter.

Which one of the two methods to choose depends on the nature of your data. If your data consists of measurements, where each variable is logically independent of other variables – typically size measurements of artefacts – you should probably use correlation coefficients. If on the other hand there is a logical dependence between variables you should probably use covariance coefficients. Typically, this could be measures of composition, whether an alloy composition or an artefact composition – but then you might well choose to use a CA in stead.

I do not believe there is a clear cut answer to what type of coefficients you should prefer in different situations, but as in archaeology we use PCA as an explorative tool the most sensible thing to do is probably to use both to see what creates the most meaningful results.

Using weights in a PCA

Weights used in a PCA are numbers between 0 and 1 including both. Weights can be applied to both objects and variables. A weight of 0 implies that an object or variable should be excluded from the analysis, while a weight of 1 implies that an object or variable should be included in the analysis. If weights are not present all objects and variables will be included. A number between 0 and 1 is used as a multiplication factor that changes all values of the associated object or variable.

The obvious use of weights as multiplication factors in a PCA is to remove unwanted side-effects of the measurements established through the variables. Take pottery for instance. If you have taken a series of measurements of different parts of the pots to discern morphological differences, you will soon find out that the size of the pots will be the dominating result of the analysis and not their shape. To counter this problem you will have to get rid of the size factor. One way to do this is to calculate the approximate volume of each pot using the measurements and then use the volumes to establish a weight factor (largest volume/ volume of pot). Or you may simply take the most dominating size variable – say pot height and use this to create a weight factor (largest height/ height of pot). You would then of course have to exclude pot height from the analysis.

Checking normality and applying transformations of variables

If you are using correlation coefficients 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 transformations based on either *Log(10)* or *Arc Sin* are implemented. Log transformations will only work with positive number. You will not be able to use transformations, if you have chosen the covariance matrix input.

There are two buttons on the PCA page called *Run normality check* and *Run PCA analysis*. If you are using covariance coefficients the *Run PCA analysis* is the only button enabled, as normality is not an issue. If, on the other hand, correlation coefficients are used the *Run normality check* is the only button enabled. As a minimum you have to produce the information on normality before you run the analysis, even if you do not use it.

When you press *Run normality check* a new worksheet is added to your workbook. It is named the same as your data worksheet with the addition *normality(PCA)*. If the resulting name is longer than 32 chars the first part is abbreviated (that is *normality(PCA)* will always be present). The first part of the worksheet shows the data as analysed. Below this a set of summary statistics are printed (Fig. 7). The statistics shown are: sum, mean, standard deviation, skewness and kurtosis. Skewness and kurtosis are of main interest here.

Skewness is a measure that shows the degree of asymmetry around the mean value of a variable. It attains zero for perfect symmetry, has a growing positive value with a growing asymmetric tail towards higher values, and has a growing negative value with a growing asymmetric tail towards lower 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 (narrow) distribution and negative values a too low (broad) 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.

Following the summary statistics is a section showing what the normality distributions will be for each variable using either Log transformation or ArcSin transformation. Looking at the example in Fig. 7 it can be seen that normality can be improved using transformations. Whether, the use of transformations will improve the results is another matter. They will certainly change them, but whether they become more interpretable is entirely up to you to decide. As we use PCA in an exploratory way there are no rules saying that you must transform your data. You decide.

Sum	1298,040	857,790	127,305	376,170	117,740	22,991	150,990	25,830	785,040	212,050	19,312
Mean	8,484	5,606	0,832	2,459	0,770	0,150	0,987	0,169	5,131	1,386	0,126
Standard Deviation	3,997	2,584	0,406	1,251	0,470	0,063	1,109	0,209	1,970	0,802	0,120
Skewness	0,980	1,106	1,325	0,515	0,629	0,092	2,355	2,852	0,357	1,502	2,168
Kurtosis	1,013	1,644	2,516	-0,345	0,138	0,224	8,326	13,882	0,701	3,161	6,094
Normality distributions following the use of transformations											
Log Transformation											
Skewness	-0,387	-0,100	0,159	-0,613	-0,868	-3,707	-0,957	-0,198	-1,001	0,265	-0,565
Kurtosis	1,067	0,215	-0,330	0,149	0,382	21,854	2,315	-0,919	2,389	-0,195	1,095
ArcSin Transformation											
Skewness	0,743	0,510	-2,980	0,622	-1,577	0,353	-1,091	-0,104	0,704	-0,631	0,067
Kurtosis	0,159	-0,368	14,165	0,086	1,133	-1,340	1,032	-1,004	1,253	-0,086	-1,546
Selection boxes for choice of transformations											
Transformation choice:	No transform	No transform	No transform	No transform	No transform	No transform	No transform	No transform	No transform	No transform	No transform

Figure 7. Part of the normality(PCA) sheet showing summary statistics for each variable, normality distributions with different types of transformations, and selection boxes for choosing transformation types.

To apply a transformation you use the last section in the *normality(PCA)* sheet (Fig. 7). It consists of one row of cells. In each of these are written *No transform*. When you activate one of these cells a small down arrow will appear at the end of the cell. If you click this arrow a drop-down box appears where you can select the transformation type you want (Fig. 8).

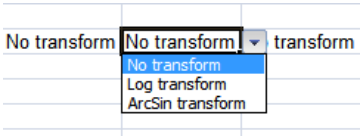


Figure 8. Selecting transformation type in the *normality(PCA)* sheet.

When you have activated the *Run normality check* button to create the *normality(PCA)* page the button will be disabled and instead the *Run PCA analysis* will be enabled.

Running the analysis

Pressing the *Run PCA analysis* will run a PCA of your data and subsequently add a new worksheet to your workbook. It is named the same as your data worksheet with the addition *statistics(PCA)*. If the resulting name is longer than 32 chars the first part is abbreviated (that is *statistics(PCA)* will always be present). The worksheet will contain the results of the analysis in tabular form.

The first section of the sheet shows the matrix of correlations coefficient used as the starting point for calculations (Fig. 9) – or the matrix of covariance coefficients, if this was the starting point of the analysis.

For correlation coefficients (Pearson's r) the coefficients are calculated as:

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

For covariance coefficients the coefficients are calculated as:

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

Correlation matrix											
	Belly height	1,000									
	Belly width	0,250	1,000								
	Belly curvature	-0,032	0,400	1,000							
	Shoulder heigth	-0,472	-0,259	-0,030	1,000						
	Shoulder width	-0,638	-0,247	-0,077	0,713	1,000					
	Shoulder curvature	-0,055	0,074	0,100	0,049	0,085	1,000				
	Neck base heigth	-0,105	0,167	-0,025	-0,148	0,046	-0,096	1,000			
	Neck base width	0,145	-0,098	-0,050	-0,089	-0,030	-0,222	0,448	1,000		
	Neck height	-0,694	-0,414	-0,133	0,011	0,225	-0,069	-0,051	-0,048	1,000	
	Neck width	-0,247	-0,322	0,052	-0,061	0,006	-0,016	-0,097	0,172	0,421	1,000
	Neck curvature	0,130	0,095	-0,170	-0,109	-0,063	-0,045	-0,017	0,031	-0,119	-0,035
		Belly height	Belly width	Belly curvature	Shoulder heigth	Shoulder width	Shoulder curvature	Neck base heigth	Neck base width	Neck height	Neck width

Figure 9. Correlation coefficient matrix used as a starting point for the PCA.

The second section shows the *Eigen values* for the calculated *Principal components* (Fig. 10). The eigenvalues reflects the amount of information that is associated with the individual principal components or eigenvectors. Knowing the total score of the eigenvalues we can calculate how large a part of the total information each component covers. Consequently, we know how large a percentage of the total information is explained by the first, the second, etc. principal components.

Eigenvalue information			
		1. Principal component	2. Principal component	3. Principal component
	Eigen values	2,778	1,701	1,442
	Explanation %	25,251	15,463	13,111
	Cumulative explanation %	25,251	40,715	53,826

Figure 10. Information on eigenvalues and explanation percentages for a PCA.

The third section shows the variable loadings (Fig. 11). That is the coordinates of the variables on the individual principal components (axes).

If you have used transformations of the variables to obtain normality, the type of transformation will be noted in brackets following the variable names.

Variable loadings				
		1. Principal component	2. Principal component	3. Principal component
	Belly height	0,516	-0,055	0,096
	Belly width (log)	0,360	0,285	-0,195
	Belly curvature (ArcSin)	0,107	0,266	-0,025
	Shoulder heigth	-0,383	0,344	-0,282
	Shoulder width	-0,451	0,232	-0,350
	Shoulder curvature (ArcSin)	-0,022	0,360	0,148
	Neck base heigth (log)	0,060	-0,281	-0,608
	Neck base width (log)	0,056	-0,493	-0,433
	Neck height (log)	-0,415	-0,275	0,257
	Neck width (log)	-0,221	-0,372	0,325
	Neck curvature (ArcSin)	0,119	-0,083	-0,001

Figure 11. Information on variable loadings for a PCA.

The fourth and final section shows object scores (Fig. 12). That is the coordinates of the objects on the individual principal components (axes).

Object scores

	1. Principal component	2. Principal component	3. Principal component
1	0,078	0,126	0,184
2	0,014	0,126	0,055
3	0,146	-0,135	-0,005
4	0,079	-0,120	0,046
6	0,096	-0,139	-0,042
7	0,207	-0,060	-0,069
8	0,158	-0,175	-0,046
9	-0,039	-0,089	0,225
10	0,136	-0,128	0,031
11	0,296	-0,078	-0,127
12	0,165	0,080	0,025
13	0,093	-0,121	0,177
14	0,076	-0,133	-0,125
15	0,330	-0,105	-0,078
16	0,220	0,064	0,039
17	0,333	0,088	0,141

Figure 12. Information on object scores for a PCA.

Running a CA

Fig. 13 shows the CA page of the main control form. This page is enabled if the validation of data has shown that a CA can be run with the data provided. Otherwise it will be disabled. The page provides controls for setting the number of Eigen vectors to be calculated and the precision with which they should be calculated (se the chapter on *Calculating Eigen vectors – number and precision*). When the analysis has been completed, the number of iterations will be displayed at the bottom of the form.

Values acceptable for CA analysis

All values in the dataset must be positive integers including zero. No special characters including (-), (.) and (,) are allowed, or stated differently negative values and real numbers are not allowed. For practical reasons blank cells are allowed. They will be interpreted as zeroes and converted to such on import. It is only in CA that blank cells are read as zeroes. In PCA they are not allowed and in MS they are treated differently. Remember that Excel as a rule will align everything it reads as a number to the right and everything else to the left. Anything you see aligned to the left in your data in the worksheet will trigger an error.

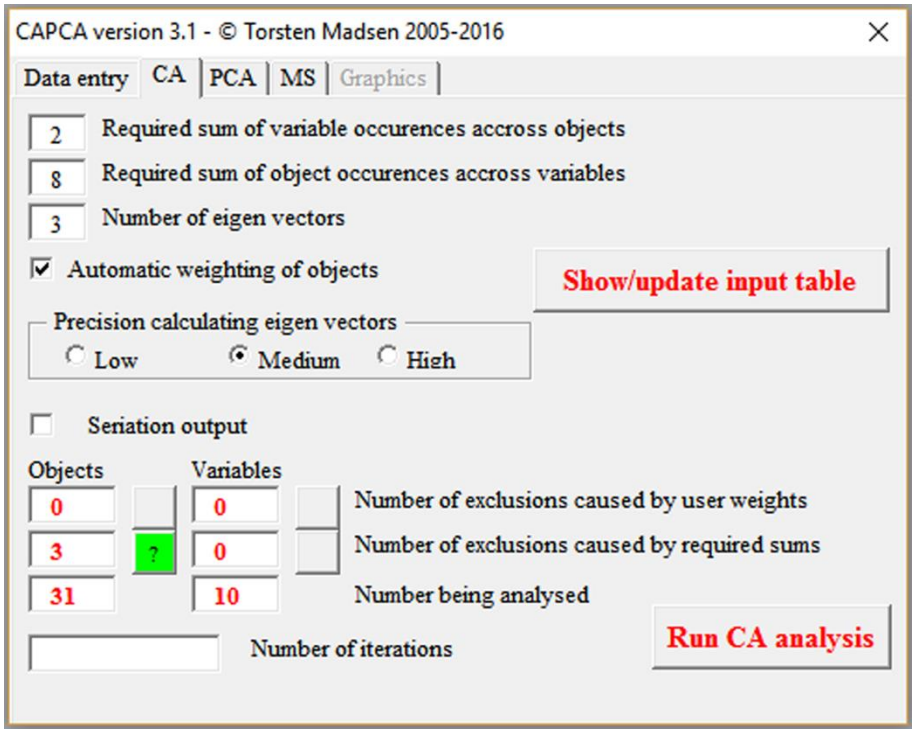


Figure 13. The CA page of the CAPCA control form

Using weights in a CA

The use of weights in a CA is much more complicated than in a PCA and of far greater value. First of all, weighting is actually in use regardless of whether you have supplied weighs or not. The reason for this is that there is a required sum of objects across variables and variables across objects that must be observed. For computational reasons the minimum must be 1. If you supply either objects or variables with a sum of 0 they will automatically be excluded. You can set the minimum number of sums for both objects and variables in the form. The default value for both is 2. Objects and variables with a lower sum than the one supplied will be removed.

You can also supply you own weights of course. These are numbers between 0 and 1 both included. Weights can be applied to both objects and variables. A weight of 0 implies that the object or the variable should be excluded from the analysis, while a weight of 1 implies that an object or variable should be included in the analysis. A number between 0 and 1 is used as a multiplication factor that changes all values of the associated object or variable.

Six info-fields inform you on how many objects and variables are being analysed and how many (if any) objects and variables have been excluded by user weights and/or by required sums. If exclusions have occurred a green button next to the number becomes available. Pressing this will open a message box giving the names of the objects and/or variables excluded.

You should note that the user weights in combination with required sums, or indeed required sums alone may start a chain reaction of exclusions, where further objects or variables falls for the sums requirements as numbers dwindle because of previous exclusions.

The use of weighting (multiplication) factors between 0 and 1is closely associated with the concept of mass and inertia of objects and variables in a CA (see below). In short the result of a CA is heavily influenced by the absolute size (mass) of an object or a variable as well as how much they divert in composition from the average (inertia). Both mass and inertia may result in what is known as outliers that dominates the result and obscures the structure in the rest of the material on the first few principal components. To lessen the effect of mass and inertia, weights are an effective tool.

There is one classical situation where mass can become a major problem. If you are analysing settlement units based on their content of artefact types, you will almost always have a situation where the units differ considerably with respect to number of artefacts. It can be a result of the actual size of the units, but more often than not, it is the result of how large a part that has been excavated. The traditional way to counter this problem in archaeology has been to calculate percentages, but in principle percentages should only be calculated for sums above 100, and under no circumstances below 50. To use a direct percentage calculation in a CA is therefore a bad idea, as a CA safely can handle sums of objects and variables below 100 in this type of material. An alternative approach is to weight objects with a sum above 100 down to this figure, but leave the objects with a sum below 100 as they are. What the minimum sum should be in this type of study is open to individual evaluation, but 10 seem to be a safe size and in many cases you can probably go lower.

For situations like the one outlined above I have included an option box for automatic weighting of objects that will weight down all objects with a sum above 100 down to a sum of 100. You can then combine this with a required sum of objects to set up a suitable analysis. For sparse matrices like counts of types in graves or presence/absence registrations this facility is of no use of course.

Matrix showing values analysed and values before weighting

Indicate objects and/or variables that has been removed due to a weight of 0 or due to minimum sums.
Values displayed are those used in the analysis.
Where values have been changed or removed due to a weighting factor, the original values are shown in brackets.

	Whipped cord	Twisted cord	Ordinary stab-and-drag	Broad stab-and-drag	Furrows	Edge stab	Circular stab	Chisel stab	Finger stab	Plastic ornament	Object sums	Original object sums
Bønnerup	0 (1)	5 (15)	22 (68)	20 (61)	0	0	16 (50)	36 (109)	0	0	99	304
Toftum A1	7 (17)	4 (9)	2 (4)	0	13 (31)	4 (10)	45 (110)	10 (25)	12 (29)	5 (12)	102	247
Toftum A31	7 (12)	3 (5)	0	0	29 (48)	14 (23)	31 (50)	7 (12)	2 (4)	6 (9)	99	163
Stengade 2	0	4 (6)	0	0	1	1	41 (64)	13 (20)	26 (40)	15 (23)	101	155
Lindebjerg 1	0	13 (15)	3	0	3 (4)	6 (7)	54 (63)	13 (15)	3	5 (6)	100	116
Havnelev	0	4	3	0	5 (6)	7 (8)	48 (54)	19 (21)	5 (6)	10 (11)	101	113
Bistoft	3	0	0	0	74	2	17	0	4	0	100	100
Toftum A46	0	5	0	0	23	13	22	1	0	2	66	66
Lindebjerg 2	0	24	1	0	2	5	17	12	3	1	65	65
Toftum A6	5	6	2	0	5	8	18	6	5	4	59	59
Värby	0	0	0	0	0	19	17	2	8	7	53	53
Mosegården	0	29	5	1	1	0	5	1	2	1	45	45
Stengade 1	0	9	0	0	0	2	10	8	11	5	45	45
Tolstrup 3	0	13	3	0	6	0	4	10	1	5	42	42
Østergårds mark	0	9	2	4	0	0	10	7	0	1	33	33
Store Valby	0	0	0	0	0	7	18	1	2	4	32	32
Yssel Bakke	1	4	2	0	0	3	7	4	2	2	25	25
Svaleklint	1	1	1	0	1	2	14	1	3	0	24	24
Moesgård skovmølle	0	15	2	0	0	0	4	0	0	0	21	21
Knardrup galgebakke	9	0	0	0	2	2	4	2	1	0	20	20
Rustrup	0	12	1	0	2	0	4	0	0	1	20	20
Tolstrup 2	0	1	1	0	0	0	5	0	3	6	16	16
Mølbjerg	0	3	0	0	5	0	1	4	0	1	14	14
Gug	0	6	1	2	0	0	0	3	0	1	13	13
Verup	3	0	0	0	1	3	4	1	1	0	13	13
Voejl	0	3	3	0	0	0	0	0	0	6	12	12
Virum	3	0	0	0	0	2	6	0	0	0	11	11
Stilling	0	0	0	0	0	0	6	1	1	1	9	9
Gilhøj	0	2	0	0	1	0	2	1	2	0	8	8
Lendrup	0	0	4	1	0	0	0	3	0	0	8	8
Slotsbjergby	0	1	2	0	0	1	1	2	0	1	8	8
Taarup	0 (1)	0 (3)	0 (1)	0 (1)	0	0	0	0 (1)	0	0	0	7
Olsbjerg	0	0	0	0	0	0	0 (4)	0	0 (2)	0	0	6
Ryumgård	0 (2)	0 (1)	0	0	0 (2)	0	0 (1)	0	0	0	0	6
Variable sums	39	176	60	28	174	101	431	168	97	90		
Original variable sum:	58	201	109	70	215	118	592	273	133	110		

Figure 14. Table of data before and after weighting.

Viewing input data before running the CA

Because of the often complex alterations to your data following the use of weighting, it can be useful to audit the resulting input data before you run the analysis. If you press *Show/update input data* a new worksheet is added to your workbook. It is named the same as your data worksheet with the addition *tables(CA)*. If the resulting name is longer than 32 chars the first part is abbreviated (that is *tables(CA)* will always be present).

The first part of the worksheet shows a table of the data to be analysed combined with the original data in brackets if changes has occurred due to weighting (Fig. 14). If the weighting has resulted in the removal of variables and/or objects these are marked out in red. Below the table, two lines show the sum of variables after weighting and the original sums. To the right of the table two columns show the sum of objects after weighting and the original sums.

The table in Fig. 14 shows the scenario outlined above with an automatic weighting combined with minimum sum of objects for a number of settlement units.

The second part of the worksheet shows a table of over- and under-representations in the dataset under the assumption that the dataset is unstructured (Fig. 15). Unstructured is here defined as a matrix of randomised occurrences based on the actual object and variable sums. More specifically, if we take the sums of objects and the sums of variables as given, we can calculate a table of occurrences as a randomisation based on these sums. This is done as follows:

$$R(a,b) = \frac{\sum_{i=1}^{n} Ma_i}{N} * \frac{\sum_{i=1}^{m} Mb_i}{N} * N$$

Where *R* is the randomized matrix, *M* is the data matrix, *a* is a given row, *b* is a given column, *m* is the number of rows, *n* is the number of columns, and *N* is the grand total of values in the data matrix.

Matrix shoving over-representations (positive) and under-representations (negative) of occurenses.
The values reflect discrepancies from a probability matrix of occurences based on object and variable sums.
Due to rounding errors the numbers given are only approximate and the grand total of the matrix is seldom zero."

	Whipped cord	Twisted cord	Ordinary stab-and-drag	Broad stab-and-drag	Furrows	Edge stab	Circular stab	Chisel stab	Finger stab	Plastic ornament	Inertia % of objects
Bønnerup	-2	-6	13	13	-9	-5	-11	17	-5	-5	17
Toftum A1	3	-7	-2	-2	0	-2	9	-2	3	-1	1
Toftum A31	3	-7	-3	-2	12	5	-1	-4	-3	-1	3
Stengade 2	-2	-7	-3	-2	-9	-5	7	0	14	6	5
Lindebjerg 1	-2	0	-1	-2	-7	-1	17	0	-3	-1	2
Havnelev	-2	-7	-1	-2	-6	0	12	5	-1	2	2
Bistoft	0	-9	-3	-2	45	-4	-11	-9	-2	-5	18
Toftum A46	-1	-3	-2	-1	11	6	1	-5	-3	-2	3
Lindebjerg 2	-1	11	-1	-1	-5	0	-3	3	-1	-2	2
Toftum A6	2	-1	0	-1	-2	3	0	-1	1	0	1
Värby	-1	-5	-2	-1	-5	11	0	-3	3	3	5
Mosegården	-1	17	2	0	-3	-2	-7	-3	-1	-1	6
Stengade 1	-1	2	-1	-1	-4	-1	-3	2	6	2	2
Tolstrup 3	-1	6	1	-1	0	-2	-7	4	-1	2	2
Østergårds mark	-1	3	0	2	-3	-2	0	2	-2	-1	2
Store Valby	-1	-3	-1	0	-3	3	6	-2	0	1	2
Yssel Bakke	0	1	1	0	-2	1	-1	1	0	0	0
Svaleklint	0	-2	0	0	-2	0	5	-1	1	-1	1
Moesgård skovmølle	0	9	1	0	-2	-1	-2	-2	-1	-1	4
Knardrup galgebakke	6	-2	-1	0	0	0	-2	0	0	-1	7
Rustrup	0	7	0	0	0	-1	-2	-2	-1	0	2
Tolstrup 2	0	-1	0	0	-2	-1	0	-1	1	4	2
Mølbjerg	0	1	0	0	2	-1	-3	2	-1	0	1
Gug	0	3	0	1	-1	-1	-3	1	-1	0	2
Verup	2	-1	0	0	0	1	0	0	0	-1	1
Voejl	0	1	2	0	-1	-1	-3	-1	-1	4	3
Virum	2	-1	0	0	-1	1	2	-1	-1	-1	2
Stilling	0	-1	0	0	-1	0	2	0	0	0	0
Gilhøj	0	1	0	0	0	0	0	0	1	0	0
Lendrup	0	-1	3	1	-1	0	-2	1	0	0	3
Slotsbjergby	0	0	1	0	-1	0	-1	1	0	0	1
Inertia % of variables	11	17	9	11	23	7	5	6	6	6	

Figure 15. over- and under-representations of occurrences in the data set.

What is shown (Fig. 15) is the differences between the actual values in the data matrix and the values in the randomised matrix. This representation of data is very useful as it is the actual starting point for a CA. It immediately gives you an impression of where the main discrepancies are between what is and what should be expected – discrepancies that will structure the results of the analysis.

Along the margins of the table the inertia percentages are given. What is shown here is simply how large a part of the total set of discrepancies is associated with the individual objects and with the individual variables. The higher an inertia percentage the higher the influence of an object or a variable on the result of the analysis becomes.

Statistic output from running the CA
When you press *Run CA* a new worksheet is added to your workbook. It is named the same as your data worksheet with the addition *statistics(CA)*. If the resulting name is longer than 32 chars the first part is abbreviated (that is *statistics(CA)* will always be present). The above mentioned worksheet *tables(CA)*will be updated, or if it is not there already, it will be created.

Eigen values and explanation percentages of eigen vectors

	Axis 1	Axis 2	Axis 3
Eigen values *10	5	3	2
Explanation %	33	21	17
Cumulative explanation %	33	54	71

Figure 16. Table of Eigen values and explanation percentages.

The first section of the statistics provides you with the Eigen values of the calculated principal components or axes, their explanation %, and the corresponding cumulative explanation percentages (Fig. 16).

Inertia and contributions of variables

			Absolute contributions			Absolute contributions weighted by eigenvalues		
	Mass%	Inertia%	Axis 1	Axis 2	Axis 3	Axis 1	Axis 2	Axis 3
Whipped cord	3	11	2	0	3	1	0	1
Twisted cord	13	17	14	3	75	6	1	18
Ordinary stab-and-drag	4	9	8	5	3	4	1	1
Broad stab-and-drag	2	11	10	10	14	5	3	3
Furrows	13	23	60	66	1	27	19	0
Edge stab	7	7	1	3	0	1	1	0
Circular stab	31	5	0	4	0	0	1	0
Chisel stab	12	6	3	0	3	1	0	1
Finger stab	7	6	0	5	0	0	2	0
Plastic ornament	7	6	0	3	0	0	1	0

Figure 17. Table of mass, inertia and absolute contributions of variables.

The next two sections of diagnostic data are termed *Mass, inertia and contributions of variables* (Fig. 17) and *Mass, inertia and contributions of objects* (Fig. 18) respectively. The first column marked *Mass %* shows the sum of counts 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 next block of columns 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^{th} object, f_{ik} is the coordinate value for the i^{th} object on the k^{th} principal axis, n is the number of objects, and i' is an object.

Inertia and contributions of objects

	Mass%	Inertia%	Absolute contributions Axis 1	Axis 2	Axis 3	Absolute contributions weighted by eigenvalues Axis 1	Axis 2	Axis 3
Bønnerup	7	17	36	25	67	16	7	16
Toftum A1	7	1	0	0	0	0	0	0
Toftum A31	7	3	2	0	0	1	0	0
Stengade 2	7	5	0	6	0	0	2	0
Lindebjerg 1	7	2	0	0	0	0	0	0
Havnelev	7	2	0	0	0	0	0	0
Bistoft	7	18	52	58	0	24	17	0
Toftum A46	5	3	2	0	0	1	0	0
Lindebjerg 2	5	2	0	0	2	0	0	0
Toftum A6	4	1	0	0	0	0	0	0
Värby	4	5	0	4	0	0	1	0
Mosegården	3	6	3	1	16	1	0	4
Stengade 1	3	2	0	0	0	0	0	0
Tolstrup 3	3	2	0	0	1	0	0	0
Østergårds mark	2	2	1	0	0	0	0	0
Store Valby	2	2	0	1	0	0	0	0
Yssel Bakke	2	0	0	0	0	0	0	0
Svaleklint	2	1	0	0	0	0	0	0
Moesgård skovmølle	2	4	1	0	6	0	0	1
Knardrup galgebakke	1	7	1	0	2	0	0	1
Rustrup	1	2	0	0	3	0	0	1
Tolstrup 2	1	2	0	0	0	0	0	0
Mølbjerg	1	1	0	0	0	0	0	0
Gug	1	2	0	0	0	0	0	0
Verup	1	1	0	0	0	0	0	0
Voejl	1	3	0	0	0	0	0	0
Virum	1	2	0	0	0	0	0	0
Stilling	1	0	0	0	0	0	0	0
Gilhøj	1	0	0	0	0	0	0	0
Lendrup	1	3	1	1	1	0	0	0
Slotsbjergby	1	1	0	0	0	0	0	0
Taarup	1	1	0	0	0	0	0	0

Figure 18. Table of mass, inertia and absolute contributions of objects.

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^{th} variable, g_{jk} is the coordinate value for the j^{th} variable on the k^{th} principal axis, n is the number of variables, and j' is a variable.

The last block of columns shows the absolute contributions to the principal axes weighted by the Eigen values associated with each axes. This is done to make the absolute contributions to the axes more comparable across the axes.

The next two sections of diagnostic data are termed *Variable coordinates (loadings)* (Fig. 19) and *Object coordinates (scores)* (Fig. 20) respectively. As with PCA the loadings and scores are coordinates of the variables and objects on the principal components, but contrary to PCA it is possible in a correspondence analysis to scale the values in such a way that they meaningfully can be represented in the same coordinate system.

Variable coordinates (loadings)	1. Principal component	2. Principal component	3. Principal component
Whipped cord	1,05	-0,19	1,17
Twisted cord	-1,02	0,47	-2,22
Ordinary stab-and-drag	-1,82	1,31	1,04
Broad stab-and-drag	-2,58	2,37	2,84
Furrows	1,81	1,77	-0,17
Edge stab	0,66	-0,88	0,34
Circular stab	0,17	-0,60	0,19
Chisel stab	-0,82	0,24	0,79
Finger stab	0,16	-1,30	-0,01
Plastic ornament	-0,12	-1,11	-0,17

Figure 19. Table of variable coordinates.

Object coordinates (scores)	1. Principal component	2. Principal component	3. Principal component
Bønnerup	-1,84	1,43	2,08
Toftum A1	0,40	-0,44	0,33
Toftum A31	0,98	0,25	0,25
Stengade 2	-0,05	-1,27	0,14
Lindebjerg 1	-0,13	-0,52	-0,10
Havnelev	-0,02	-0,62	0,38
Bistoft	2,11	2,08	-0,11
Toftum A46	1,08	0,46	-0,19
Lindebjerg 2	-0,59	-0,02	-1,21
Toftum A6	0,21	-0,44	0,13
Värby	0,40	-1,54	0,38
Mosegården	-1,30	0,73	-2,50
Stengade 1	-0,38	-0,88	-0,55
Tolstrup 3	-0,56	0,60	-0,93
Østergårds mark	-1,23	0,61	0,05
Store Valby	0,31	-1,36	0,37
Yssel Bakke	-0,40	-0,47	-0,04
Svaleklint	0,21	-0,80	0,33
Moesgård skovmølle	-1,29	0,64	-2,98
Knardrup galgebakke	1,01	-0,29	1,36
Rustrup	-0,73	0,64	-2,60
Tolstrup 2	-0,21	-1,35	-0,17
Mølbjerg	0,29	1,25	-0,64
Gug	-1,79	1,20	-0,69
Verup	0,80	-0,69	0,93
Voejl	-1,14	-0,20	-0,78
Virum	0,75	-0,99	0,99
Stilling	0,04	-1,18	0,40
Gilhøj	-0,07	-0,20	-0,89
Lendrup	-2,29	1,91	2,41
Slotsbjergby	-1,04	0,22	0,46
Taarup	-1,53	1,35	-0,24

Figure 20. Table of objects coordinates.

Using the seriation option

The CA page has a further option named *seriation output*. If you check this option a seriation of the analysed data will be carried out. BUT NOTE! This does not mean that the data can be meaningfully seriated. You should consult the paper Multivariate data analysis using PCA, CA and MS in CAPCA available together with this manual on the use of CA for seriation purposes.

In short, a material that can be seriated will show an arched layout of both variables and objects in a plot of the two first principal components. A second degree polynomial can be used to describe this layout. If the polynomial fits the points well, a seriation of the data is meaningful and a sorted version of the data constituting a seriation can be created.

Seriated table of data based on CA

	Broad stab-and-drag	Ordinary stab-and-drag	Twisted cord	Chisel stab	Plastic ornament	Finger stab	Circular stab	Edge stab	Whipped cord	Furrows
Lendrup	1	4	0	3	0	0	0	0	0	0
Bønnerup	20	22	5	36	0	0	16	0	0	0
Taarup	1	1	3	1	0	0	0	0	1	0
Gug	2	1	6	3	1	0	0	0	0	0
Mosegården	1	5	29	1	1	2	5	0	0	1
Moesgård skovmølle	0	2	15	0	0	0	4	0	0	0
Østergårds mark	4	2	9	7	1	0	10	0	0	0
Rustrup	0	1	12	0	1	0	4	0	0	2
Tolstrup 3	0	3	13	10	5	1	4	0	0	6
Slotsbjergby	0	2	1	2	1	0	1	1	0	0
Voejl	0	3	3	0	6	0	0	0	0	0
Lindebjerg 2	0	1	24	12	1	3	17	5	0	2
Yssel Bakke	0	2	4	4	2	2	7	3	1	0
Stengade 1	0	0	9	8	5	11	10	2	0	0
Gilhøj	0	0	2	1	0	2	2	0	0	1
Lindebjerg 1	0	3	13	13	5	3	54	6	0	3
Tolstrup 2	0	1	1	0	6	3	5	0	0	0
Havnelev	0	3	4	19	10	5	48	7	0	5
Stengade 2	0	0	4	13	15	26	41	1	0	1
Stilling	0	0	0	1	1	1	6	0	0	0
Svaleklint	0	1	1	1	0	3	14	2	1	1
Store Valby	0	0	0	1	4	2	18	7	0	0
Värby	0	0	0	2	7	8	17	19	0	0
Toftum A6	0	2	6	6	4	5	18	8	5	5
Toftum A1	0	2	4	10	5	12	45	4	7	13
Virum	0	0	0	0	0	0	6	2	3	0
Verup	0	0	0	1	0	1	4	3	3	1
Knardrup galgebakke	0	0	0	2	0	1	4	2	9	2
Toftum A31	0	0	3	7	6	2	31	14	7	29
Toftum A46	0	0	5	1	2	0	22	13	0	23
Mølbjerg	0	0	3	4	1	0	1	0	0	5
Bistoft	0	0	0	0	0	4	17	2	3	74

The polynomial used as basis for seriation of objects is: 0,5650X^2 -0,0812X + -0,5894
With a Pearson R^2 of 0,6485 for goodness of fit.
The polynomial used as basis for seriation of variables is: 0,5613X^2 + 0,0658X + -0,7061
With a Pearson R^2 of 0,8473 for goodness of fit.

Figure 21. Content of the seriation(CA) worksheet

The polynomials to be used (one for the variables and one for the objects) are found through regression, and the over all goodness of fit for the variables and objects to these polynomials are evaluated through Pearsons R^2. This figure will be 1 for a perfect fit (that is all points lies precisely on the curve described by the polynomials) and 0 for a total randomness of the points in relation to the curve.

If you have checked *Seriation output*, a new worksheet is added to your workbook. It is named the same as your data worksheet with the addition *seriation(CA)*. If the resulting name is longer than 32 chars the first part is abbreviated (that is *seriation(CA)* will always be present).

The worksheet presents you with a sorted version of your data (Fig. 21). The sorting is based on the polynomials calculated for the variables and objects respectively. Each variable is projected orthogonally onto the curve for the variables, and each object is projected orthogonally onto the curve for the objects. The resulting order of the variables along the curve for variables is used to

sort the columns of the data matrix, and the resulting order of the variables along the curve for objects is used to sort the rows of the data matrix.

Below the sorted matrix you will find the equation of the polynomial used to sort the objects and the equation of the polynomial used to sort the variables together with the respective values of the Pearson R^2 for goodness of fit. In the actual example the Pearson R^2 is not impressive telling us that the data is ill fitted for seriation.

Note! You should never use the seriation worksheet alone. You should always consult the graphical output, where you can also get the polynomials drawn through the plots of variables and objects. A bad seriation is often created by ill fitted objects or variables that profitably can be left out. These are most easily identified through the graphics output.

Running MS

Metric scaling operates from a matrix of similarity coefficients between a set of objects. It thus deviates from a CA and a PCA by not analysing objects against variables directly. A similarity coefficient matrix is a square matrix that holds the objects in both rows and columns. The cell values show how similar two objects are to each other. This leads to three fundamental characteristic of the matrix:

- The diagonal cells of the matrix, where objects are compared to themselves, must all have the same value – a value that represents identity.
- The values in the off-diagonal cells must be smaller or equal to the value of the diagonal cells. They cannot be larger.
- Since two objects are compared with each other twice in cells on each sides of the diagonal, the values of the matrix are always symmetric around the diagonal.

The strength of a MS is that you are not bound by a particular scheme of describing objects by variables (as measures for a PCA, or counts for a CA). You can create your own system for comparing objects. The weakness, on the other hand, is that the connection to the descriptive elements you use is broken, and it is not possible to see how individual variables have influenced the results.

There are three different ways in which you can enter data for a MS in CAPCA.

Entering data as a similarity matrix

To enter a similarity matrix directly you must first check the similarity/distance matrix option on the main data entry page. To CAPCA this indicates that data entered should meet certain criteria. As outlined above the data matrix must be square and symmetric around the diagonal. All values on the diagonal must be the same and they must be larger or equal to the off diagonal values. You can use positive integer or real values, but not negative values. Object names must be entered, while anything entered as variable names is ignored. Weights are not applicable, and if entered they are ignored.

It is customary, but not mandatory to scale the similarity coefficients so that the diagonal values become 1 and the off diagonal values lies between 1 and 0. For practical reasons CAPCA will automatically rescale the coefficients so that the diagonal values become 1.

Entering data as a distance matrix

A distance matrix is the inverse of a similarity matrix so to speak. Think of a table of distances between cities in an automobile atlas to get the idea of what it is. The main difference from a similarity matrix is of course that the smallest values, namely 0, must be on the diagonal while off diagonal values must be larger.

To enter a distance matrix directly you must first check the similarity/distance matrix option on the main data entry page. To CAPCA this indicates that data entered should meet certain criteria. As with the similarity coefficients the data matrix must be square and symmetric around the diagonal. All values on the diagonal must be 0, while the off diagonal values must be larger than 0 (or 0). You can use positive integer or real values, but not negative values. Object names must be entered, while anything entered as variable names is ignored. Weights are not applicable, and if entered they are ignored.

As MS always analyses a similarity coefficient matrix, the distance matrix is transformed into a similarity coefficient matrix before analysis.

Entering data as objects and variables

You can enter data for MS as objects and variables and let CAPCA calculate the similarity coefficients, but you then have to accept the method of calculation used in CAPCA, and you have to comply with some basic rules for the formatting of variables. The method used is known as Gower's general coefficient of similarity (Gower 1971). The method recognises three types of variables: Continuous quantitative variables; Categorical dichotomous variables, Categorical qualitative variables.

Missing values

In contrast to PCA and CA, MS accepts missing values. This is so because the variables are used to calculate the similarity coefficients exclusively, and are not involved with the results presented. If a value is missing it is simply omitted from the calculation of the similarity coefficient. If many missing values are present it will influence the validity of the similarity coefficients, of course. To signal a missing value in a cell you simply enter a question mark in the cell. It will work with both continuous quantitative variables, categorical dichotomous variables and categorical qualitative variables.

Continuous quantitative variables

Valid values are all kinds of measurements, percentages and counts (be careful if you use raw counts). They can be either integers or real numbers, negative or positive, and include zero. Apart from question marks signalling missing values the only special characters allowed are (-) (.) and (,). They are interpreted according to the national setup of your computer (e.g. -2.455,23 in a Danish setup compared -2,455.23 in an English setup). Blank cells are allowed. They will be interpreted as zeroes. As text is allowed in categorical qualitative variables (see below), there can be no rigid control of the content of the variables you supply. If something in a variable cannot be interpreted as a number, the variable will automatically be interpreted as a categorical qualitative variable, which certainly may have unforeseen consequences. If a variable exclusively contain values of 1 and 0 it will automatically interpreted as a categorical presence/absence variable.

Categorical dichotomous variables

Categorical dichotomous variables are simple recordings of whether a variable (a trait) is present in an object or not. 1 and 0 is used to record these two possibilities. Blank cells are allowed. They will be interpreted as zeroes.

Values accepted for Categorical qualitative variables

Categorical qualitative variables are multistate variables. Typically it would be classifications like Romanesque, Gothic and Barok. You write the classes directly in the cells as text strings (In contrast to version 3.0 of CAPCA where you should code them as negative integers). As with the other types of variables you can use question marks to signal missing values. Blank cells are not allowed.

Calculating the similarity coefficients

To calculate a similarity coefficient between two objects they are compared across their variables, one at a time. 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. If one or either of the variables holds a question mark for missing data, the comparison is not valid and neither *Validity* nor *Scores* are incremented.

For categorical 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 categorical qualitative variables *Scores* is incremented with 1 if both objects display the same element and is not incremented if they differ.

For continuous quantitative 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.

Creating a similarity coefficient matrix
If you have provided either a similarity or distance coefficient matrix, the MS page of the CAPCA main form will show you how many objects the matrix contains, while the variables box shows *na* (not applicable) (Fig. 22). Below this is a button named *Check similarity/distance matrix*. You have to press this before you can press the *Run MS analysis*.

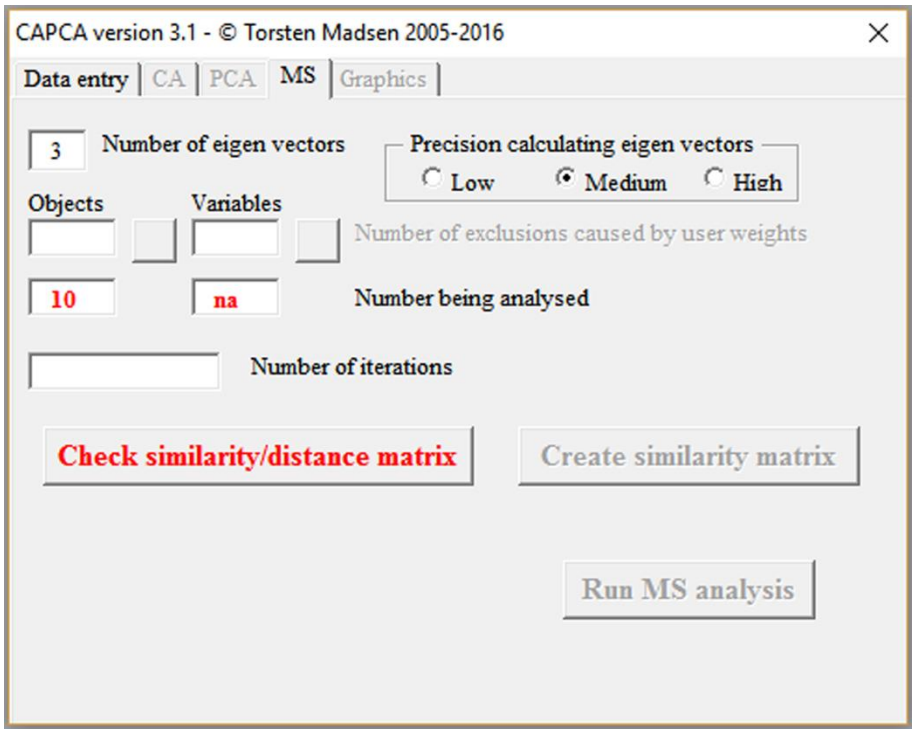


Figure 22. MS page of the CAPCA main form as it appears when you have entered a similarity or distance coefficient matrix.

Activating the *Check similarity/distance matrix* button will add a new worksheet to your workbook. It is named the same as your data worksheet with the addition *similarity(MS)*. If the resulting name is longer than 32 chars the first part is abbreviated (that is *similarity(MS)* will always be present).
The worksheet contains two sections (Fig. 23). The first shows the similarity or distance matrix exactly as it was entered. The second shows the similarity matrix as it will be analysed. In the example in Fig. 23 a distance matrix was entered that subsequently was transformed into a similarity coefficient matrix. If you enter a similarity coefficient matrix with the same format as used in CAPCA the two matrices will be identical of course.

Coefficient matrix showing values as read from the worksheet

	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

Coefficient matrix showing values as they will be analysed

	Amsterdam	Berlin	Brussel	Budapest	Genève	Luxembourg	Paris	Praha	Warszawa	Wien
Amsterdam	1	0,59	0,87	0,12	0,44	0,76	0,69	0,47	0,24	0,28
Berlin	0,59	1	0,52	0,47	0,33	0,52	0,35	0,79	0,63	0,61
Brussel	0,87	0,52	1	0,15	0,56	0,87	0,81	0,45	0,17	0,31
Budapest	0,12	0,47	0,15	1	0,2	0,26	0,07	0,68	0,58	0,85
Genève	0,44	0,33	0,56	0,2	1	0,68	0,69	0,43	0,03	0,36
Luxembourg	0,76	0,52	0,87	0,26	0,68	1	0,78	0,55	0,2	0,42
Paris	0,69	0,35	0,81	0,07	0,69	0,78	1	0,36	0	0,23
Praha	0,47	0,79	0,45	0,68	0,43	0,55	0,36	1	0,62	0,82
Warszawa	0,24	0,63	0,17	0,58	0,03	0,2	0	0,62	1	0,57
Wien	0,28	0,61	0,31	0,85	0,36	0,42	0,23	0,82	0,57	1

Figure 23. The content of the similarity(MS) worksheet when a distance matrix have been entered.

If you have supplied data as objects and variables the MS page of the CAPCA main form will show you how many objects and variables that are included in the analysis (Fig. 24). Further you will find information on how many objects and variables that are excluded due to weighting if any. Exclusion will only occur as a result of user weights. Below the information boxes is a button named *Create similarity matrix*. You have to press this before you can press the *Run MS analysis*.

The worksheet with the addition *similarity(MS)* mentioned above will also be created, when you activate this button, and again it will contain two sections (Fig. 25). The first section shows the data that you have entered, but in most cases not in the same format. The variables have been analysed by CAPCA and split into three groups named “Quantitative”, “Dichotomous” and “Qualitative”. Regardless of the order in which you submitted the variables they will always be shown in the order outlined above and with a fixed format. It is important to check that there is an agreement between the way CAPCA has classified the variables and the way you intended them to be classified.

The second section shows the similarity coefficient matrix that has been calculated from the variables using Gowers method of calculation. You may note that there is no longer a traceable connection between the coefficients and the variables, which of course is an inherent weakness in using similarity coefficients.

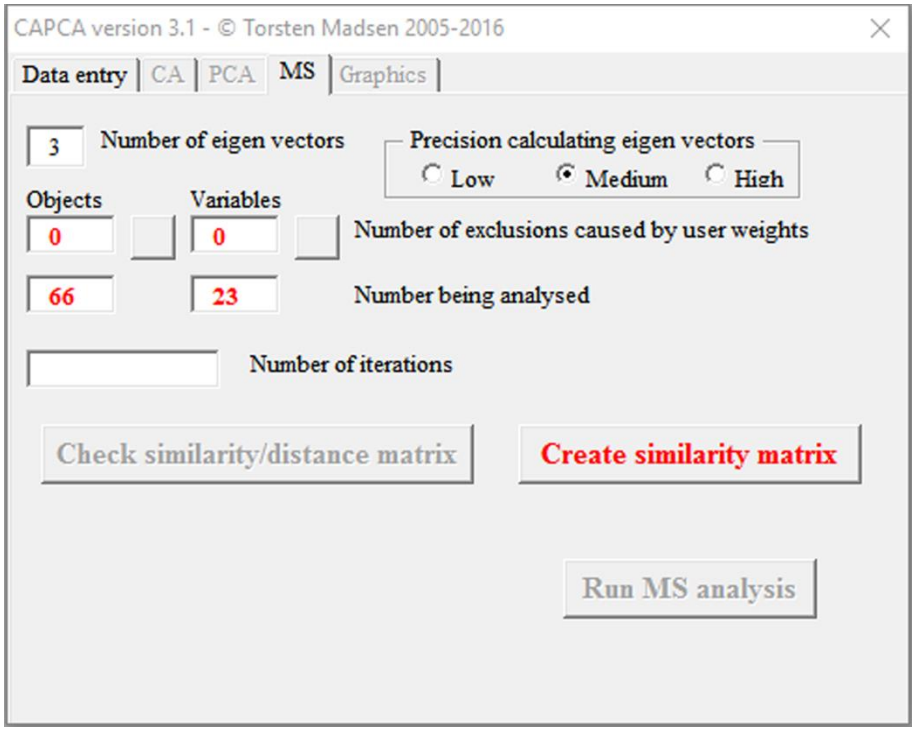
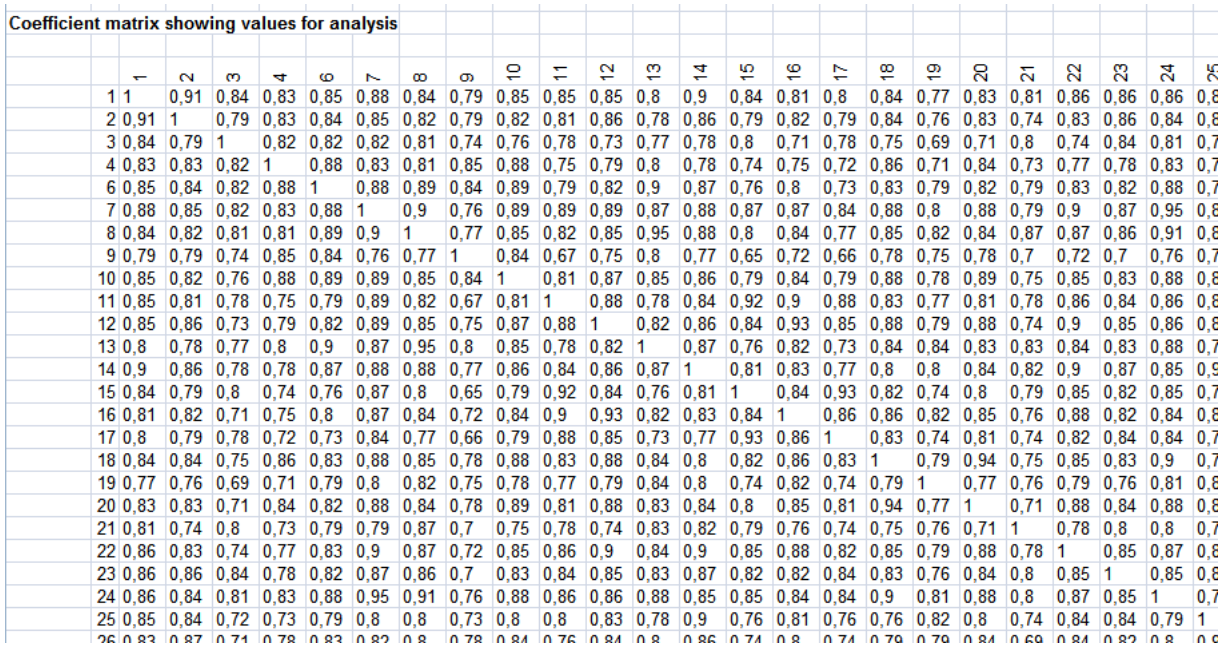


Figure 24. MS page of the CAPCA main form as it appears when you have entered data as objects and variables.

Table showing variables as read acording to variable type for calculating a similarity matrix

NB! quantitative values are presented with two decimals, regardless of input format.

	Quantitative													Dichotomous										
	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	RT Fingernails	RT Irregular stabs	RT Oblong cuts	RT Triangular stabs	RT short strokes	RT Stab-and-drag	RT Multiple pointed stab	RT Two-ply cord	RT Obligue stabs	RT vertical stabs	RT Incised lines
1	1,09	3,31	2,01	0,31	0,88	0,14	0,01	0,08	0,01	1,39	0,76	0	1	0	0	0	0	0	0	0	0	0	0
2	1,05	2,84	2,05	0,35	1,17	0,16	0,02	0,24	0,03	1,42	0,66	0,02	0	1	0	0	0	0	0	0	0	0	0
3	0,73	3,3	2,34	0,35	0,4	0,07	0,11	0	0	1,85	0,86	0,01	0	0	0	0	0	0	0	0	0	0	0
4	1,08	2,99	1,93	0,39	0,71	0,08	0,06	0,16	0	1,65	0,86	0,08	0	0	0	0	0	0	0	0	0	0	0
6	1,38	3,1	1,94	0,28	0,79	0,07	0,06	0	0	1,7	0,63	0,05	0	1	0	0	0	0	0	0	0	0	0
7	1,12	3,71	2,17	0,32	0,68	0,09	0,04	0	0	1,26	0,56	0,04	0	0	0	0	0	0	0	0	0	0	0
8	1,25	3,82	1,83	0,25	0,52	0,15	0,1	0	0	1,42	0,61	0,04	0	0	0	0	0	0	0	0	0	0	0
9	1,27	2,88	1,5	0,23	0,94	0,04	0,01	0,08	0,03	2,11	0,85	0,07	0	0	0	0	0	0	0	0	0	0	0
10	1,15	3,51	1,85	0,31	0,76	0,05	0	0,16	0	1,64	0,51	0,07	0	0	0	0	0	0	0	0	0	0	0
11	1,14	4,12	2,14	0,37	0,61	0,07	0,02	0	0	1,26	0,26	0	0	0	0	0	0	0	0	0	0	0	0
12	1,21	3,74	1,88	0,37	0,86	0,12	0,02	0,24	0,02	1,12	0,36	0,03	1	0	0	0	0	0	0	0	0	0	0
13	1,35	3,84	1,55	0,24	0,58	0,17	0,1	0,01	0	1,52	0,59	0,05	0	0	0	0	0	0	0	0	0	0	0
14	1,29	3,46	1,76	0,29	0,98	0,23	0,04	0	0	1,41	0,53	0,01	0	0	1	0	0	0	0	0	0	0	0
15	1,05	4,32	2,3	0,39	0,25	0,08	0,01	0	0	1,12	0,47	0	0	0	0	0	0	0	0	0	0	0	0
16	1,36	4,06	1,82	0,33	0,63	0,09	0,02	0,25	0,03	1,11	0,27	0,03	0	0	0	0	0	0	0	0	0	0	0
17	0,85	4,42	2,25	0,39	0,35	0,03	0	0,11	0,04	1,12	0,43	0,01	0	0	0	0	0	0	0	0	0	0	0
18	0,94	3,99	1,92	0,35	0,59	0,12	0,02	0,15	0,03	1,18	0,64	0,07	0	0	0	0	0	0	0	0	0	0	0
19	1,51	3,96	1,37	0,11	0,61	0,08	0,02	0,13	0,11	1,49	0,59	0,02	0	0	0	0	0	0	0	0	0	0	0
20	1,01	3,81	1,88	0,33	0,77	0,22	0,02	0,11	0,04	1,19	0,53	0,08	0	0	0	0	0	0	0	0	0	0	0
21	1,26	4,01	1,78	0,25	0,22	0,07	0,24	0	0	1,41	0,74	0,01	0	0	0	0	0	0	0	0	0	0	0
22	1,14	4,01	1,89	0,31	0,85	0,26	0,04	0,03	0	0,94	0,5	0,03	0	0	0	0	0	0	0	0	0	0	0
23	0,86	3,6	2,06	0,35	0,73	0,18	0,13	0,1	0,03	1,46	0,47	0,02	0	0	0	0	0	0	0	0	0	0	0
24	0,98	3,83	2,17	0,29	0,51	0,09	0,04	0,13	0	1,32	0,6	0,05	0	0	0	0	0	0	0	0	0	0	0
25	1,11	3,33	1,83	0,23	1,09	0,26	0,04	0,13	0,11	1,45	0,41	0,01	0	0	0	0	0	0	0	0	0	0	0
26	1,05	3,06	1,73	0,3	1,27	0,27	0,02	0,11	0,08	1,61	0,46	0,04	0	1	0	0	0	0	0	0	0	0	0



Coefficient matrix showing values for analysis

	1	2	3	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	1	0,91	0,84	0,83	0,85	0,88	0,84	0,79	0,85	0,85	0,85	0,8	0,9	0,84	0,81	0,8	0,84	0,77	0,83	0,81	0,86	0,86	0,86	0,8
2	0,91	1	0,79	0,83	0,84	0,85	0,82	0,79	0,82	0,81	0,86	0,78	0,86	0,79	0,82	0,79	0,84	0,76	0,83	0,74	0,83	0,86	0,84	0,8
3	0,84	0,79	1	0,82	0,82	0,82	0,81	0,74	0,76	0,78	0,73	0,77	0,78	0,8	0,71	0,78	0,75	0,69	0,71	0,8	0,74	0,84	0,81	0,7
4	0,83	0,83	0,82	1	0,88	0,83	0,81	0,85	0,88	0,75	0,79	0,8	0,78	0,74	0,75	0,72	0,86	0,71	0,84	0,73	0,77	0,78	0,83	0,7
6	0,85	0,84	0,82	0,88	1	0,88	0,89	0,84	0,89	0,79	0,82	0,9	0,87	0,76	0,8	0,73	0,83	0,79	0,82	0,79	0,83	0,82	0,88	0,7
7	0,88	0,85	0,82	0,83	0,88	1	0,9	0,76	0,89	0,89	0,89	0,87	0,88	0,87	0,87	0,84	0,88	0,8	0,88	0,79	0,9	0,87	0,95	0,8
8	0,84	0,82	0,81	0,81	0,89	0,9	1	0,77	0,85	0,82	0,85	0,95	0,88	0,8	0,84	0,77	0,85	0,82	0,84	0,87	0,87	0,86	0,91	0,8
9	0,79	0,79	0,74	0,85	0,84	0,76	0,77	1	0,84	0,67	0,75	0,8	0,77	0,65	0,72	0,66	0,78	0,75	0,78	0,7	0,72	0,7	0,76	0,7
10	0,85	0,82	0,76	0,88	0,89	0,89	0,85	0,84	1	0,81	0,87	0,85	0,86	0,79	0,84	0,79	0,88	0,78	0,89	0,75	0,85	0,83	0,88	0,8
11	0,85	0,81	0,78	0,75	0,79	0,89	0,82	0,67	0,81	1	0,88	0,78	0,84	0,92	0,9	0,88	0,83	0,77	0,81	0,78	0,86	0,84	0,86	0,8
12	0,85	0,86	0,73	0,79	0,82	0,89	0,85	0,75	0,87	0,88	1	0,82	0,86	0,84	0,93	0,85	0,88	0,79	0,88	0,74	0,9	0,85	0,86	0,8
13	0,8	0,78	0,77	0,8	0,9	0,87	0,95	0,8	0,85	0,78	0,82	1	0,87	0,76	0,82	0,73	0,84	0,84	0,83	0,83	0,84	0,83	0,88	0,7
14	0,9	0,86	0,78	0,78	0,87	0,88	0,88	0,77	0,86	0,84	0,86	0,87	1	0,81	0,83	0,77	0,8	0,8	0,84	0,82	0,9	0,87	0,85	0,9
15	0,84	0,79	0,8	0,74	0,76	0,87	0,8	0,65	0,79	0,92	0,84	0,76	0,81	1	0,84	0,93	0,82	0,74	0,8	0,79	0,85	0,82	0,85	0,7
16	0,81	0,82	0,71	0,75	0,8	0,87	0,84	0,72	0,84	0,9	0,93	0,82	0,83	0,84	1	0,86	0,86	0,82	0,85	0,76	0,88	0,82	0,84	0,8
17	0,8	0,79	0,78	0,72	0,73	0,84	0,77	0,66	0,79	0,88	0,85	0,73	0,77	0,93	0,86	1	0,83	0,74	0,81	0,74	0,82	0,84	0,84	0,7
18	0,84	0,84	0,75	0,86	0,83	0,88	0,85	0,78	0,88	0,83	0,88	0,84	0,8	0,82	0,86	0,83	1	0,79	0,94	0,75	0,85	0,83	0,9	0,7
19	0,77	0,76	0,69	0,71	0,79	0,8	0,82	0,75	0,78	0,77	0,79	0,84	0,8	0,74	0,82	0,74	0,79	1	0,77	0,76	0,79	0,76	0,81	0,8
20	0,83	0,83	0,71	0,84	0,82	0,88	0,84	0,78	0,89	0,81	0,88	0,83	0,84	0,8	0,85	0,81	0,94	0,77	1	0,71	0,88	0,84	0,88	0,8
21	0,81	0,74	0,8	0,73	0,79	0,79	0,87	0,7	0,75	0,78	0,74	0,83	0,82	0,79	0,76	0,74	0,75	0,76	0,71	1	0,78	0,8	0,8	0,7
22	0,86	0,83	0,74	0,77	0,83	0,9	0,87	0,72	0,85	0,86	0,9	0,84	0,9	0,85	0,88	0,82	0,85	0,79	0,88	0,78	1	0,85	0,87	0,8
23	0,86	0,86	0,84	0,78	0,82	0,87	0,86	0,7	0,83	0,84	0,85	0,83	0,87	0,82	0,82	0,84	0,83	0,76	0,84	0,8	0,85	1	0,85	0,8
24	0,86	0,84	0,81	0,83	0,88	0,95	0,91	0,76	0,88	0,86	0,86	0,88	0,85	0,85	0,84	0,84	0,9	0,81	0,88	0,8	0,87	0,85	1	0,7
25	0,85	0,84	0,72	0,73	0,79	0,8	0,8	0,73	0,8	0,8	0,83	0,78	0,9	0,76	0,81	0,76	0,76	0,82	0,8	0,74	0,84	0,84	0,79	1
26	0,83	0,87	0,71	0,78	0,83	0,82	0,8	0,78	0,84	0,76	0,84	0,8	0,86	0,74	0,8	0,74	0,79	0,79	0,84	0,69	0,84	0,82	0,8	0,8

Figure 25. The content of the similarity(MS) worksheet when data was entered as objects and variables. In the example the variables entered were partly quantitative and partly dichotomous. Note that the printout has been cropped.

Graphical presentation

For all three types of analysis, the best way to view the result is through a plot of objects and variables based on the principal axes/components. When you have successfully completed an analysis the Graphics page of the main control form will become available Fig. 26).

The page features two ways of depicting the result. One is through two-dimensional scatter plots based on the principal axes as analysed. The other is through a metric scaling of a number of the principal axes, presented as a two-dimensional scatter plot. The latter is a new feature in version 3.1 compared to version 3.0.

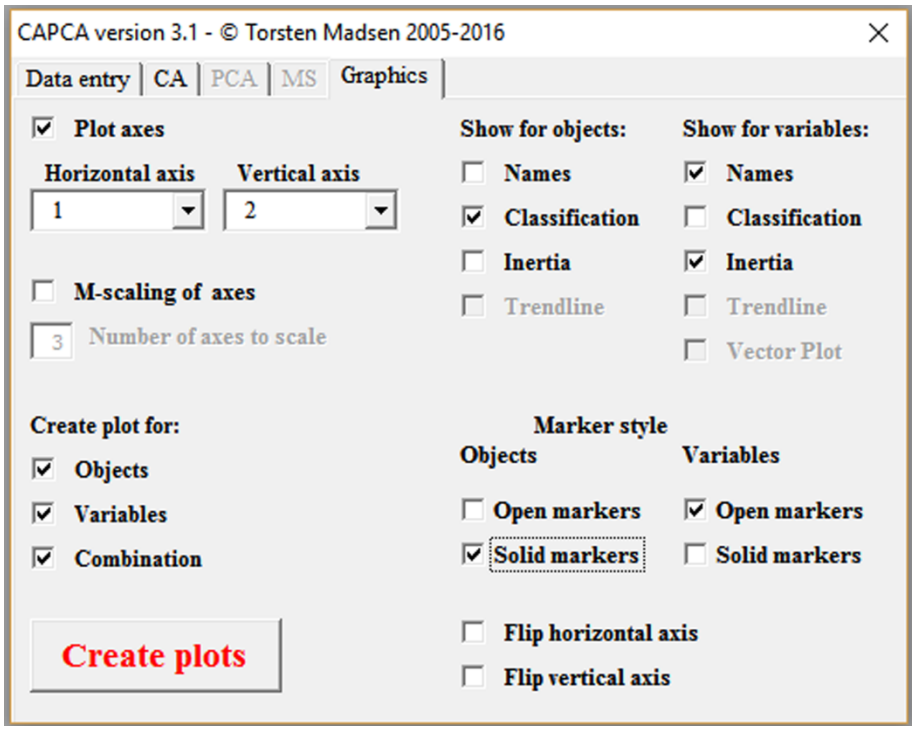


Figure 26. The Graphics page of the main control form with the option Plot axes checked.

Plotting principal axes

To plot individual axes against each other you check the option control "Plot axes". There is a number of settings to control the content and layout of the graphical presentation of the axes. As all plots are scatter charts with both a horizontal and a vertical axis, the first you should do is to select the two principal axes to appear on the horizontal and vertical axes. The default is the 1st and 2nd principal axes on the horizontal and vertical axes respectively, but using the drop down boxes you can choose any combination and order of the calculated principal axes.

The second thing to choose is the kind of content the plots should have. Three types of plots are available: Object plots (available for CA, PCA and MS), Variable plots (available for CA and PCA) and Combination plots (available for CA). When you check *Objects* a number of settings for objects will become available. When you check *Variables* a number of settings for variables will become available. When you check *Combinations* settings for both objects and variables will become available.

For both objects and variables you can choose what kind of information should appear in the plots. If you check the option *Names* the names of the objects and variables will be shown adjacent to markers representing the objects and variables in the plot. If you check the option *Classification*

(only available if you have supplied a classification) the plots will show the classification using different types of markers for the individual classes (Fig. 27).

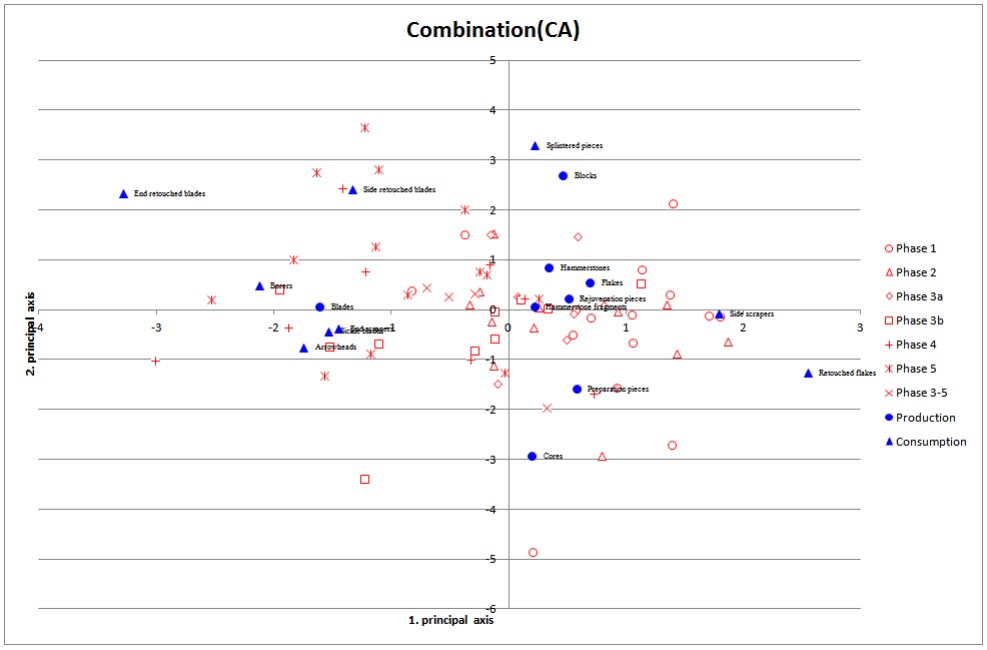


Figure 27. Combination plot from a CA with the *Classification* option checked, and with open objects markers and solid variable markers.

For CA the option *Inertia* will be available for both objects and variables. If you check this the size of the markers will vary with the magnitude of the inertia of an object or variable (Fig. 28).

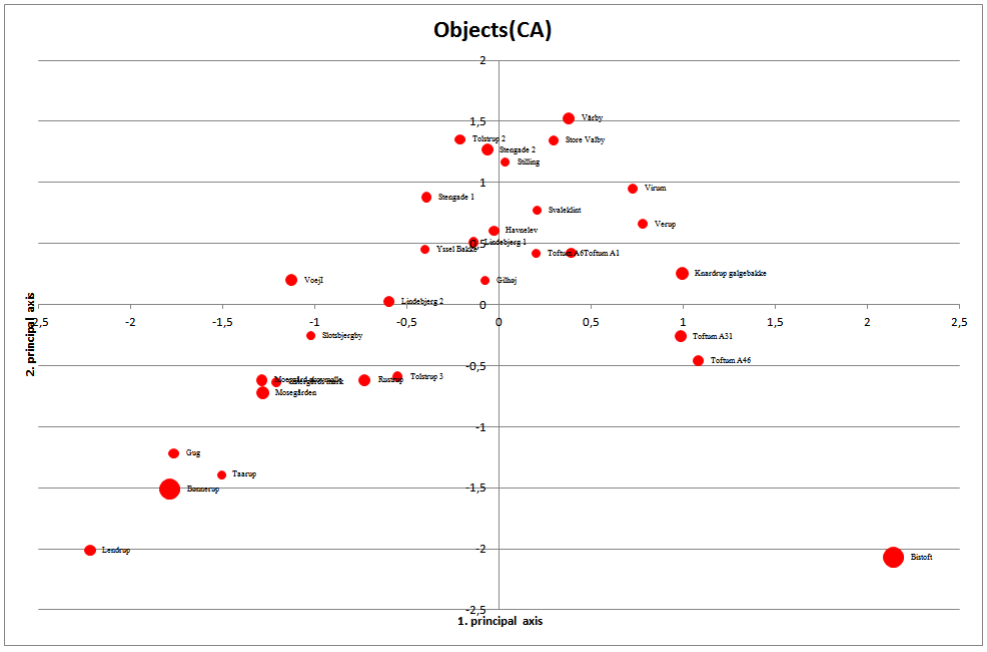


Figure 28. Objects plot from a CA with the *Inertia* option checked

If you have run a CA with the option *Seriation* checked, the option *Trendline* will be available for both objects and variables. A second degree polynomial fitted to the points in the plot is displayed together with its equations and Pearsons R for goodness of fit (Fig.29). If you check *Trendline* the option *Inertia* will be disabled.

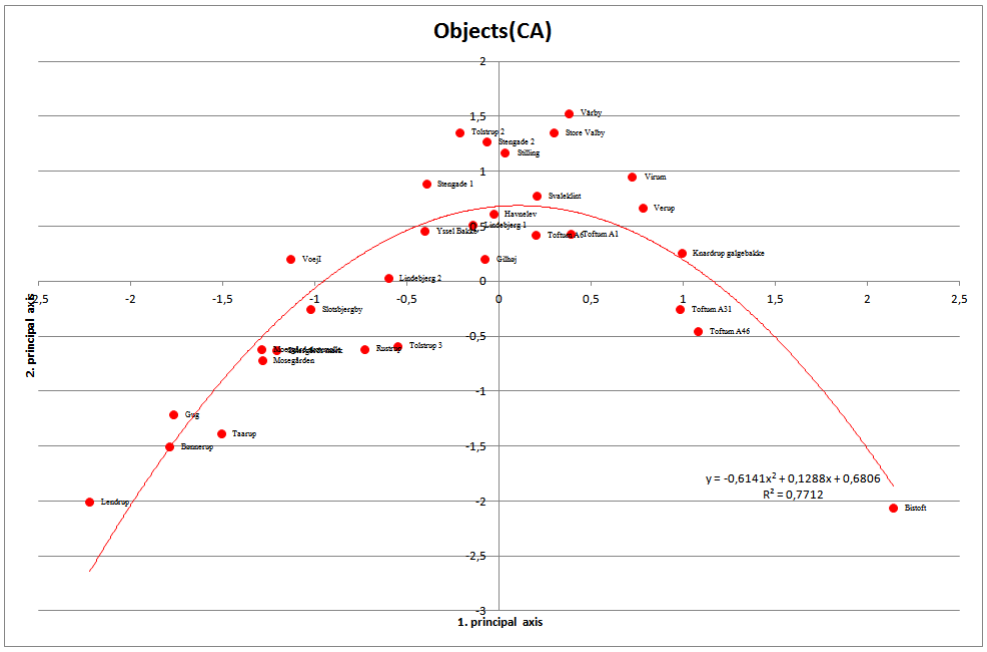


Figure 29. Objects plot from a CA with the *Seriation* and *Trendline* options checked.

If you have run a PCA, the option Vector Plot will be available for variables. This will substitute the normal points plot with vectors drawn from (0,0) to the position of the variables in the plot (Fig. 30).

For both objects and variables you can choose whether the plots should be shown with open markers (i.e. with only contour lines) or with solid markers. Open markers makes it easier to discern overlying points, while solid markers give more clear impressions. Using open markers you have 32 different markers to work with as a combination of shape and colour for both objects and variables. Using solid markers you have only 16 markers at your disposal. At the moment you cannot choose different schemes for marker forms and colours.

The last set of options allows you to flip the axes. As there is no natural orientation of the principal axes you can freely choose the way your plots should be orientated. It can be an advantage, for instance, in a seriation, where you probably would prefer to have the oldest material to the left and the youngest to the right. Also you might find that the orientation of the axes suddenly becomes reversed following even small changes to the input data. For easier comparison of two versions of the analysis, the possibility of flipping the axes may be a help.

When you have chosen the options you want, you press the *Create plots* button. This will add a chart for each of the plot types – objects, variables and combinations – you have chosen. They will be named the same as your data worksheet with the addition *Objects(CA), Variables(CA)* and *Combination(CA)* if you have run a CA, or otherwise PCA or MS where it applies. If the resulting name is longer than 32 chars, the first part is abbreviated.

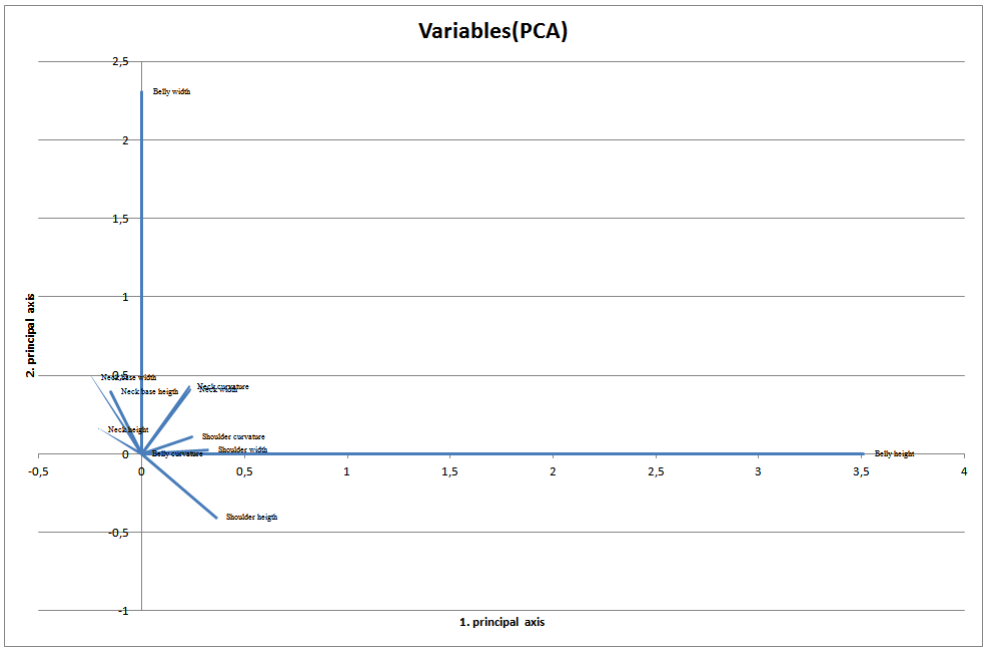


Figure 30. Variable plot from a PCA with the *Vector plo*t option checked.

Metric scaling of axes

You may consider the result of a CA, a PCA or a MS as a scatterplot of points in a multidimensional space. When we want to see and understand the results of the analysis, we have to look at the individual axes constituting this space individually, two together in a scatterplot or three together in a three-dimensional plot. The latter solution is not available in Excel (at least not in a true sense), but other programs exist, where you can create interactive three-dimensional plots.

One way to get to grips with the content of more than two axes at a time is to analyse formally the content of the axes. Each axis represents a dimension in an orthogonal space, meaning that it lies perpendicular to any other axis in this space. This is easy enough to understand for two axes and for three, but mindboggling to understand for four or more axes. Yet we can calculate Euclidian distances between points in such a space. We all know how to do such a calculation in two dimensions using Pythagoras' theorem:

$$d_{i,j} = \sqrt{d(x_i, x_j)^2 + d(y_i, y_j)^2}$$

That is, the distance between two points i and j equals the square root of the sum of squared distances between the two points along axis x and axis y.

It is also clear to see that in a three-dimensional space, we just add the distance between the two points on the third axis z to the equation, and even if it is difficult to visualize we can continue with the fourth axes, etc. in this way.

If we take each point in turn and calculate the distances to all of the other points, we will end up with a square distance matrix that we can analyse through a metric scaling, and that is what happens through the option *M-scaling of axes*. Before you start using this option, you should heed the following two warning notes:

The results of CA, PCA and MS analyses are structured in such a way that the first few principal axes hold the largest and most coherent parts of the information. As you move down though the axes, more and more "noise" will be present. Thus, it is by no means a good idea to include many axes in a scaling. It will only lead to blurred results.

You should not include axes in a scaling before you have studied them individually to see what kind of information influence and structure them. Very often, you will find outliers that do not influence the first two principal axes dominate the fifth, fourth or even the third principal axes. If you include an axes that is completely dominated by an outlier you are in for a bad experience.

My own experiences are that you may get good results with scaling three or even four axes, but seldom more, but I have also experienced outliers on the third principal axis that completely has ruined the result.

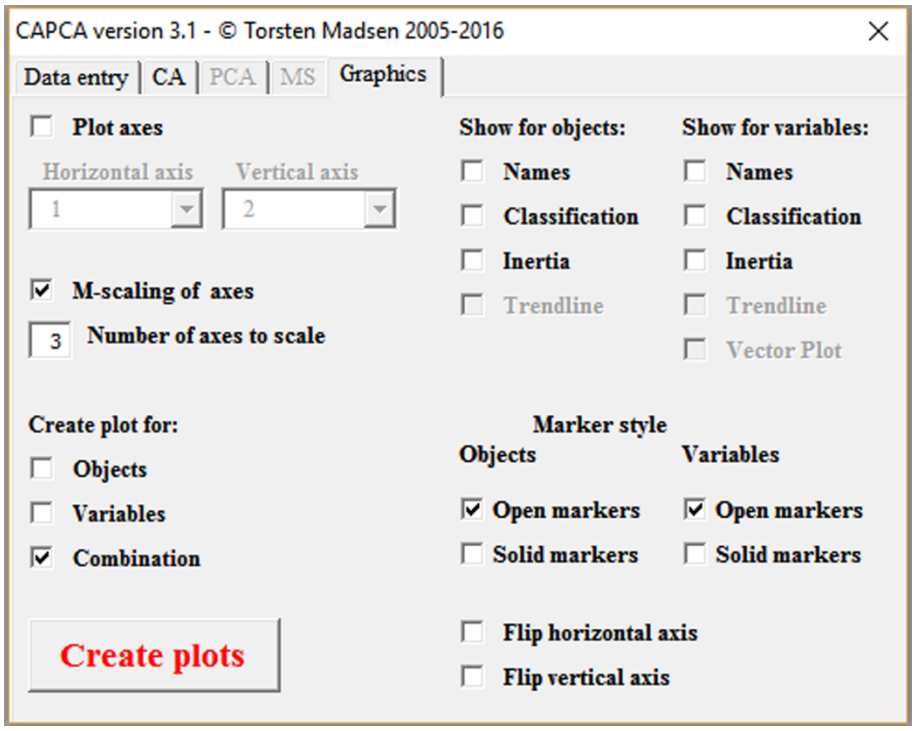


Figure 31. The Graphics page of the main control form with the option M-scaling of axes checked.

Initially, when you check the option M-scaling of axes, most controls are disabled. Those options that no matter what will stay disabled are Trendline, Vector Plot and the selection boxes for horizontal and vertical axes. Those options that will be available is governed by the same factors as with the "plot axes" option: The analysis you have run (CA, PCA or MS), the plot type (Objects, Variables or Combination) you choose, and whether you have provided classifications.

You should note that it is only possible to choose one of the plot types at a time (Objects, Variables or Combination). The reason for this is that a new analysis that directly depends on your choice of plot type has to run before the plot can be formed.

All the options to the right, works as already described above in connection with the normal plot of axes.

To the left there is one option that relates directly to scaling. In "Number of axes to scale" you can set the number of principal axes that should be included in the analysis. The default value is 3, and you should note that this implies that the first three principal axes will be analysed. If you change the value to 4 you will analyse the first four principal values and so forth.

You run the analysis (and create the plot) by pressing the *Create plots* button. When done you are returned to the page from where you originally ran an analysis (for instance CA) and the graphics page becomes disabled. You have to rerun the original analysis before you can make new

plots. The reason for this is that the metric scaling of the axes changes essential data that has to be re-established before you can make new plots.

The plot created will be named the same as your data worksheet with the addition *M-Scaling of objects, M-Scaling of variables* and *M-Scaling of combination.* If the resulting name is longer than 32 chars, the first part is abbreviated.

EDITING AND SAVING THE PLOTS

The plots are created entirely using the Microsoft Excel Chart facility. In contrast to earlier versions of CAPCA the coordinates for the plots are no longer written into a worksheet to which the plots subsequently are tied. In version 3 the plots are generated directly from data in memory and are thus no longer linked to physical data in your workbook. This has the advantage that it is much easier to save the charts with a new name.

For each set of data you analyse you can have up to three plots: an objects plot, a variable plot and a combination plot. Every time you make new plots based on the analysis (of different axes for instance) these three plots will be written over if they already exists. To preserve existent plots you have to save them under a new name. You can do that by simply changing the name of the charts.

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. You can also change the marker form, size and colour. 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 Excel 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 tif. The advantage is that the plots are stored in an easily accessible format, shared by all. One drawback is that the resolution is 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.