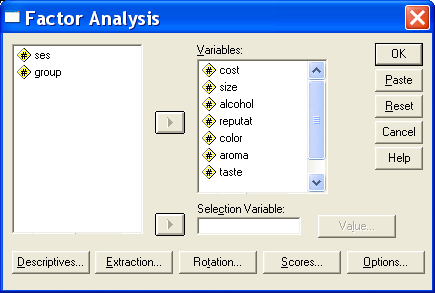
**MA4128 : EXAMPLE**

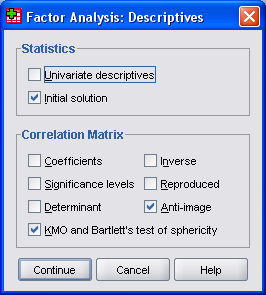
Suppose a market research company is interested in what influences a consumer’s choice behavior when e is shopping for beer. The researcher asked each of 200 consumers to rate on a scale of 0-100 how important e considers each of seven qualities when deciding whether or not to buy the six pack: low COST of the six pack, high SIZE of the bottle (volume), high percentage of ALCOHOL in the beer, the REPUTATion of the brand, the COLOR of the beer, nice AROMA of the beer, and good TASTE of the beer.

The data are in the file [FACTBEER.SAV](http://core.ecu.edu/psyc/wuenschk/SPSS/FactBeer.sav). Import that file into SPSS. On the command bar, click Analyze, Data Reduction, Factor.

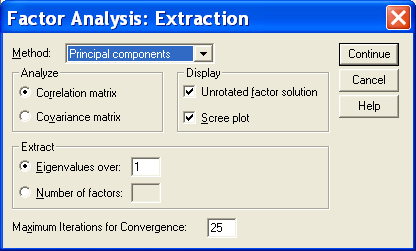
Transfer the seven variables of interest into the Variables Box:



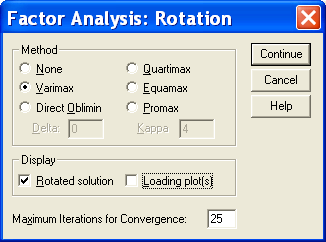
1. Click Descriptives and then check Initial Solution, Coefficients, KMO and Bartlett’s Test of Sphericity, and Anti-image. Click Continue.



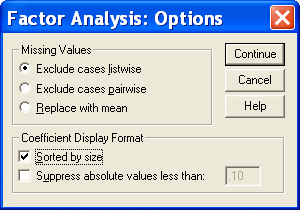
1. Click Extraction and then select Correlation Matrix, Unrotated Factor Solution, Scree Plot, and Eigenvalues Over 1. Click Continue.



1. Click Rotation. Select Varimax and Rotated Solution. Click Continue.



1. Click Options. Select Exclude Cases Listwise and Sorted By Size. Click Continue.



Click OK, and SPSS completes the Principal Components Analysis.

**Checking For Unique Variables**

Aside from the raw data matrix, the first matrix you are likely to encounter in a PCA (or Factor Analysis) is the correlation matrix. Here is the correlation matrix for our data:

| **Correlation Matrix** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | cost | size | alcohol | reputat | color | aroma | taste | ses | group |
| Correlation | cost | 1.000 | .832 | .767 | -.406 | .018 | -.046 | -.064 | -.662 | -.515 |
| size | .832 | 1.000 | .904 | -.392 | .179 | .098 | .026 | -.662 | -.572 |
| alcohol | .767 | .904 | 1.000 | -.463 | .072 | .044 | .012 | -.652 | -.636 |
| reputat | -.406 | -.392 | -.463 | 1.000 | -.372 | -.443 | -.443 | .056 | .274 |
| color | .018 | .179 | .072 | -.372 | 1.000 | .909 | .903 | .583 | .435 |
| aroma | -.046 | .098 | .044 | -.443 | .909 | 1.000 | .870 | .640 | .399 |
| taste | -.064 | .026 | .012 | -.443 | .903 | .870 | 1.000 | .681 | .500 |
| ses | -.662 | -.662 | -.652 | .056 | .583 | .640 | .681 | 1.000 | .790 |
| group | -.515 | -.572 | -.636 | .274 | .435 | .399 | .500 | .790 | 1.000 |

Unless it is just too large to grasp, you should give the correlation matrix a thorough visual inspection You are planning to use PCA to capture the essence of the correlations in this matrix. Notice that there are many medium to large correlations in this matrix, and that every variable, except reputation, has some large correlations, and reputation is moderately correlated with everything else (negatively).

There is also a statistic, Bartlett’s test of sphericity, that can be used to test the null hypothesis that our sample was randomly drawn from a population in which the correlation matrix was an identity matrix, a matrix full of zeros, except, of course, for ones on the main diagonal.

If there are any variables that are not correlated with the other variables, you might as well delete them prior to the PCA. If you are using PCA to reduce the set of variables to a smaller set of components to be used in additional analyses, you can always reintroduce the unique (not correlated with other variables) variables at that time. Alternatively, you may wish to collect more data, adding variables that you think will indeed correlate with the now unique variable, and then run the PCA on the new data set.

| **Communalities** | | |
| --- | --- | --- |
|  | Initial |
| cost | .738 |
| size | .912 |
| alcohol | .866 |
| reputat | .499 |
| color | .922 |
| aroma | .857 |
| taste | .881 |

Communalities

One may also wish to inspect the Squared Multiple Correlation coefficient (SMC or *R2* ) of each variable with all other variables. Variables with small *R2* s are unique variables, not well correlated with a linear combination of the other variables. If you conduct a principal axis factor analysis, these are found in the ***Communalities*** (Initial) table.

**Kaiser’s Measure of Sampling Adequacy**

It is undesirable to have two variables which share variance with each other but not with other variables.

Recall that the partial correlation coefficient between variables Xi and Xj is the correlation between two residuals,

 and 

A large partial correlation indicates that the variables involved share variance that is not shared by the other variables in the data set. Kaiser’s Measure of Sampling Adequacy (**MSA**) for a variable Xi is the ratio of the sum of the squared simple *r*’s between Xi and each other X to (that same sum plus the sum of the squared partial *r*’s between Xi and each other X). Recall that squared *r*’s can be thought of as variances.



Small values of *MSA* indicate that the correlations between Xi and the other variables are unique, that is, not related to the remaining variables outside each simple correlation. Kaiser has described *MSAs* above .9 as ***marvelous***, above .8 as ***meritorious***, above .7 as ***middling***, above .6 as ***mediocre***, above .5 as ***miserable***, and below .5 as ***unacceptable***.

The MSA option in ***SAS’ PROC FACTOR*** gives you a matrix of the partial correlations, the *MSA* for each variable, and an overall *MSA* computed across all variables.

Variables with small MSAs should be deleted prior to FA or the data set supplemented with additional relevant variables which one hopes will be correlated with the offending variables.

For our sample data the partial correlation matrix looks like this:

COST SIZE ALCOHOL REPUTAT COLOR AROMA TASTE

COST 1.00 .54 -.11 -.26 -.10 -.14 .11

SIZE .54 1.00 .81 .11 .50 .06 -.44

ALCOHOL -.11 .81 1.00 -.23 -.38 .06 .31

REPUTAT -.26 .11 -.23 1.00 .23 -.29 -.26

COLOR -.10 .50 -.38 .23 1.00 .57 .69

AROMA -.14 .06 .06 -.29 .57 1.00 .09

TASTE .11 -.44 .31 -.26 .69 .09 1.00

MSA .78 .55 .63 .76 .59 .80 .68

OVERALL MSA = .67

These *MSA’s* may not be marvelous, but they aren’t low enough to make me drop any variables (moreso as there is only 7 variables, already an unrealistically low number).

The SPSS output includes the overall *MSA* in the same table as the less useful Bartlett’s test of sphericity.

****

The partial correlations (each multiplied by minus 1) are found in the Anti-Image Correlation Matrix. On the main diagonal of this matrix are the *MSAs* for the individual variables.

| **Anti-image Matrices** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | cost | size | alcohol | reputat | color | aroma | taste |
| Anti-image  Correlation | cost | .779a | -.543 | .105 | .256 | .100 | .135 | -.105 |
| size | -.543 | .550a | -.806 | -.109 | -.495 | .061 | .435 |
| alcohol | .105 | -.806 | .630a | .226 | .381 | -.060 | -.310 |
| reputat | .256 | -.109 | .226 | .763a | -.231 | .287 | .257 |
| color | .100 | -.495 | .381 | -.231 | .590a | -.574 | -.693 |
| aroma | .135 | .061 | -.060 | .287 | -.574 | .801a | -.087 |
| taste | -.105 | .435 | -.310 | .257 | -.693 | -.087 | .676a |
| a. Measures of Sampling Adequacy(MSA) | | | |

**Extracting Principal Components**

We are now ready to extract principal components. We shall let the computer do most of the work, which is considerable. From *p* variables we can extract *p* components. This will involve solving *p* equations with *p* unknowns. The variance in the correlation matrix is “repackaged” into *p* eigenvalues. Each **eigenvalue** represents the amount of variance that has been captured by one component.

Each component is a linear combination of the *p* variables. The first component accounts for the largest possible amount of variance. The second component, formed from the variance remaining after that associated with the first component has been extracted, accounts for the second largest amount of variance, etc. The principal components are extracted with the restriction that they are orthogonal. Geometrically they may be viewed as dimensions in *p*-dimensional space where each dimension is perpendicular to each other dimension.

Each of the *p* variable’s variance is standardized to one. Each factor’s eigenvalue may be compared to 1 to see how much more (or less) variance it represents than does a single variable. With *p* variables there is *p* x 1 = *p* variance to distribute. The principal components extraction will produce *p* components which in the aggregate account for all of the variance in the *p* variables. That is, the sum of the *p* eigenvalues will be equal to *p*, the number of variables. The proportion of variance accounted for by one component equals its eigenvalue divided by *p*.

For our beer data, here are the eigenvalues and proportions of variance for the seven components:

****

**Deciding How Many Components to Retain**

So far, all we have done is to repackage the variance from *p* correlated variables into *p* uncorrelated components. We probably want to have fewer than *p* components. If our *p* variables do share considerable variance, several of the *p* components should have large eigenvalues and many should have small eigenvalues.

One needs to decide how many components to retain. One handy rule of thumb is to retain only components with eigenvalues of one or more.

That is, drop any component that accounts for less variance than does a single variable. Another device for deciding on the number of components to retain is the **scree test**. This is a plot with eigenvalues on the ordinate and component number on the abscissa. Scree is the rubble at the base of a sloping cliff.

In a scree plot, scree is those components that are at the bottom of the sloping plot of eigenvalues versus component number. The plot provides a visual aid for deciding at what point including additional components no longer increases the amount of variance accounted for by a nontrivial amount. Here is the scree plot produced by SPSS:

****

For our beer data, only the first two components have eigenvalues greater than 1. There is a big drop in eigenvalue between component 2 and component 3. On a scree plot, components 3 through 7 would appear as scree at the base of the cliff composed of components 1 and 2.

Together components 1 and 2 account for 85% of the total variance. We shall retain only the first two components.