**Principal Components Analysis - SPSS[[1]](#footnote-1)©**

 In principal components analysis (**PCA**) and factor analysis (**FA**) one wishes to extract from a set of *p* variables a reduced set of *m* components or factors that accounts for most of the variance in the *p* variables. In other words, we wish to reduce a set of *p* variables to a set of *m* underlying superordinate dimensions.

 These underlying factors are inferred from the correlations among the *p* variables. Each factor is estimated as a weighted sum of the *p* variables. The *ith* factor is thus

 

 One may also express each of the *p* variables as a linear combination of the *m* factors,

 

where *Uj* is the variance that is unique to variable *j*, variance that cannot be explained by any of the common factors.

**Goals of PCA and FA**

 One may do a PCA or FA simply **to reduce a set of *p* variables to *m* components or factors prior to further analyses on those *m* factors**. For example, Ossenkopp and Mazmanian (*Physiology and Behavior*, *34*: 935-941) had 19 behavioral and physiological variables from which they wished to predict a single criterion variable, physiological response to four hours of cold-restraint. They first subjected the 19 predictor variables to a FA. They extracted five factors, which were labeled Exploration, General Activity, Metabolic Rate, Behavioral Reactivity, and Autonomic Reactivity. They then computed for each subject scores on each of the five factors. That is, each subject’s set of scores on 19 variables was reduced to a set of scores on 5 factors. These five factors were then used as predictors (of the single criterion) in a stepwise multiple regression.

 One may use FA **to discover and summarize the pattern of intercorrelations among variables**. This is often called **Exploratory FA.** One simply wishes to group together (into factors) variables that are highly correlated with one another, presumably because they all are influenced by the same underlying dimension (factor). One may also then operationalize (invent a way to measure) the underlying dimension by a linear combination of the variables that contributed most heavily to the factor.

 If one has a theory regarding what basic dimensions underlie an observed event, e may engage in **Confirmatory Factor Analysis**. For example, if I believe that performance on standardized tests of academic aptitude represents the joint operation of several basically independent faculties, such as Thurstone’s Verbal Comprehension, Word Fluency, Simple Arithmetic, Spatial Ability, Associative Memory, Perceptual Speed, and General Reasoning, rather than one global intelligence factor, then I may use FA as a tool to analyze test results to see whether or not the various items on the test do fall into distinct factors that seem to represent those specific faculties.

 Psychometricians often employ FA in **test construction**. If you wish to develop a test that measures several different dimensions, each important for some reason, you first devise questions (variables) which you think will measure these dimensions. For example, you may wish to develop a test to predict how well an individual will do as a school teacher. You decide that the important dimensions are Love of Children, Love of Knowledge, Tolerance to Fiscal Poverty, Acting Ability, and Cognitive Flexibility. For each of these dimensions you write several items intended to measure the dimension. You administer the test to many people and FA the results. Hopefully many items cluster into factors representing the dimensions you intended to measure. Those items that do not so cluster are rewritten or discarded and new items are written. The new test is administered and the results factor analyzed, etc. etc. until you are pleased with the instrument. Then you go out and collect data testing which (if any) of the factors is indeed related to actual teaching performance (if you can find a valid measure thereof) or some other criterion (such as teacher’s morale).

 There are numerous other uses of FA that you may run across in the literature. For example, some researchers may investigate the differences in factor structure between groups. For example, is the factor structure of an instrument that measures socio-politico-economic dimensions the same for citizens of the U.S.A. as it is for citizens of Mainland China? Note such various applications of FA when you encounter them.

**A Simple, Contrived Example**

 Suppose I am interested in what influences a consumer’s choice behavior when e is shopping for beer. I ask 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). Bring that file into SPSS. On the command bar, click Analyze, Data Reduction, Factor. Scoot the seven variables of interest into the Variables Box:



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



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



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



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 FA is the correlation matrix. Here is the correlation matrix for our data:

 COST SIZE ALCOHOL REPUTAT COLOR AROMA TASTE

COST 1.00 .83 .77 -.41 .02 -.05 -.06

SIZE .83 1.00 .90 -.39 .18 .10 .03

ALCOHOL .77 .90 1.00 -.46 .07 .04 .01

REPUTAT -.41 -.39 -.46 1.00 -.37 -.44 -.44

COLOR .02 .18 .07 -.37 1.00 .91 .90

AROMA -.05 .10 .04 -.44 .91 1.00 .87

TASTE -.06 .03 .01 -.44 .90 .87 1.00

 Unless it is just too large to grasp, you should give the correlation matrix a good look. 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 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. I think a good ole Eyeball Test is generally more advisable, unless you just don’t want to do the PCA, someone else is trying to get you to, and you need some “official” sounding “justification” not to do it.

 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 |

 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 [Enter PROC FACTOR MSA;] 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 (especially since I have only seven variables, already an unrealistically low number).

 The SPSS output includes the overall *MSA* in the same table as the (useless) 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-imageCorrelation | 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.

 I often find it useful to try at least three different solutions, and then decide among them which packages the variance in a way most pleasing to me. Here I could try a one component, a two component, and a three component solution.

 Some folks find the methods I have outlined above too subjective. These folks will be more pleased with a statistical method for deciding how many components to retain. I shall discuss two such methods, Parallel Analysis and Velicer’s MAP test.

 **Parallel Analysis.** In this analysis one determines how many components account for more variance than do components derived from random data. One creates many random sets of data, with the constraint that each set has the same number of variables and the same number of cases as the actual set of data. One then compares the eigenvalues from an analysis of the actual data with the distribution of eigenvalues from the random sets of data. Starting with the first component, the eigenvalue from the original data is compared to the 95th percentile of the eigenvalues of the first components from the random data. If the original data produced an eigenvalue greater than that 95th percentile, then that component is retained and you move on to the second component. This process is continued until for the *nth* component, the eigenvalue from the actual data is not greater than the 95th percentile from the random data. At that point, one stops and decides to retain *n – 1* components.

 SAS, SPSS, and Matlab scripts for conducting parallel analysis can be found at <https://people.ok.ubc.ca/brioconn/nfactors/nfactors.html> . Here I illustrate use of the SPSS script. First, be sure that the file with the original data contain only the data to be used for the components analysis – any other variables need be deleted. With the data file open in SPSS, edit the script to indicate how many variables there are, how many cases there are, and how many random data sets to create. Then simply run the script. Here is output:

|  |  |
| --- | --- |
| Parallel Analysis | Actual Data |
|  **PARALLEL ANALYSIS:****Principal Components****Specifications for this Run:****Ncases 231****Nvars 7****Ndatsets 1000****Percent 95****Random Data Eigenvalues** **Root Means Prcntyle** **1.000000 1.251612 1.344920** **2.000000 1.146058 1.207526** **3.000000 1.064757 1.118462** **4.000000 .992964 1.038794** **5.000000 .926129 .973311** **6.000000 .852369 .907173** **7.000000 .766109 .830506** |  |

 Notice that for only the first two components, the eigenvalues from the actual data have values greater than those of the 95th percentile of the random data. Accordingly, a two component solution is indicated.

 The script also includes the option to replace the main diagonal of the input correlation matrix with estimates of the communalities of the variables. While this is controversial, it is my preference when the researcher intends to conduct a factor analysis rather than a components analysis.

**Velicer’s MAP Test**. With this procedure one considers how much common (shared by variables) variance remains in the data after extracting *n* components. In the first step, the first component is removed from the original correlation matrix, resulting in a matrix of partial correlations. The off-diagonal elements, when squared, represent variance shared by the variables in potential components other than the first. The mean squared off-diagonal partial correlation coefficient is computed for this step. On the second step the first two components are removed from the original correlation matrix. On the *nth* step, the first *n* components are removed from the original correlation matrix. One retains the component which has the smallest mean squared off-diagonal partial correlation (and all earlier components).

 The correlation matrix is the input for the SPSS script. It can be brought in from an external file or simply pasted into the program. Here I have done the latter:

matrix.

compute cr = {

 1.00, .83, .77, -.41, .02, -.05, -.06;

.83, 1.00, .90, -.39, .18, .10, .03;

.77, .90, 1.00, -.46, .07, .04, .01;

-.41, -.39, -.46, 1.00, -.37, -.44, -.44;

.02, .18, .07, -.37, 1.00, .91, .90;

-.05, .10, .04, -.44, .91, 1.00, .87;

-.06, .03, .01, -.44, .90, .87, 1.00 }. <snip, snip>

The output:

**Velicer's Minimum Average Partial (MAP) Test:**

**Velicer's Average Squared Correlations**

 **.000000 .266624**

 **1.000000 .440869**

 **2.000000 .129252**

 **3.000000 .170272**

 **4.000000 .331686**

 **5.000000 .486046**

 **6.000000 1.000000**

**The smallest average squared correlation is**

 **.129252**

**The number of components is**

 **2**

 Which Test to Use? With luck, both parallel analysis and the MAP test will indicate the same decision. Parallel analysis has a slight tendency to extract too many components, the MAP test too few. If the two disagree, try increasing the number of random data sets in the parallel analysis and look carefully at the two smallest squared partial correlation coefficients from the MAP test. You may end up having to use the interpretability criterion.

**Loadings, Unrotated and Rotated**

 Another matrix of interest is the **loading matrix**, also known as the **factor pattern matrix** or the **component matrix**. The entries in this matrix, loadings, are correlations between the components and the variables. Since the two components are orthogonal, these correlation coefficients are also beta weights, that is, , thus *A1* equals the number of standard deviations that *Xj* changes for each one standard deviation change in Factor 1. Here is the loading matrix for our beer data:

****

 As you can see, almost all of the variables load well on the first component, all positively excepting reputation. The second component is more interesting, with 3 large positive loadings and three large negative loadings. Component 1 seems to reflect concern for economy and quality versus reputation. Component 2 seems to reflect economy versus quality.

 Remember that each component represents an orthogonal (perpendicular) dimension. Fortunately, we retained only two dimensions, so I can plot them on paper. If we had retained more than two components, we could look at several pairwise plots (two components at a time).

 For each variable I have plotted in the vertical dimension its loading on component 1, and in the horizontal dimension its loading on component 2. Wouldn’t it be nice if I could rotate these axes so that the two dimensions passed more nearly through the two major clusters (COST, SIZE, ALCH and COLOR, AROMA, TASTE)? Imagine that the two axes are perpendicular wires joined at the origin (0,0) with a pin. I rotate them, preserving their perpendicularity, so that the one axis passes through or near the one cluster, the other through or near the other cluster. The number of degrees by which I rotate the axes is the angle **PSI**. For these data, rotating the axes -40.63 degrees has the desired effect.

 Here is the loading matrix after rotation:



**Number of Components in the Rotated Solution**

 I generally will look at the initial, unrotated, extraction and make an initial judgment regarding how many components to retain. Then I will obtain and inspect rotated solutions with that many, one less than that many, and one more than that many components. I may use a "meaningfulness" criterion to help me decide which solution to retain – if a solution leads to a component which is not well defined (has none or very few variables loading on it) or which just does not make sense, I may decide not to accept that solution.

 One can err in the direction of extracting too many components (overextraction) or too few components (underextraction). Wood, Tataryn, and Gorsuch (1996, *Psychological Methods*, *1*, 354-365) have studied the effects of under- and over-extraction in principal factor analysis with varimax rotation. They used simulation methods, sampling from populations where the true factor structure was known. They found that overextraction generally led to less error (differences between the structure of the obtained factors and that of the true factors) than did underextraction. Of course, extracting the correct number of factors is the best solution, but it might be a good strategy to lean towards overextraction to avoid the greater error found with underextraction.

 Wood et al. did find one case in which overextraction was especially problematic – the case where the true factor structure is that there is only a single factor, there are no unique variables (variables which do not share variance with others in the data set), and where the statistician extracts two factors and employs a varimax rotation (the type I used with our example data). In this case, they found that the first unrotated factor had loadings close to those of the true factor, with only low loadings on the second factor. However, after rotation, factor splitting took place – for some of the variables the obtained solution grossly underestimated their loadings on the first factor and overestimated them on the second factor. That is, the second factor was imaginary and the first factor was corrupted. Interestingly, if there were unique variables in the data set, such factor splitting was not a problem. The authors suggested that one include unique variables in the data set to avoid this potential problem. I suppose one could do this by including "filler" items on a questionnaire. The authors recommend using a random number generator to create the unique variables or manually inserting into the correlation matrix variables that have a zero correlation with all others. These unique variables can be removed for the final analysis, after determining how many factors to retain.

**Explained Variance**

 The SPSS output also gives the **variance explained by each component** after the rotation. The variance explained is equal to the **sum of squared loadings (SSL) across variables**. For component 1 that is (.762 + .742 +...+ .672) = 3.31 = its eigenvalue before rotation and (.962 + .962 +...+ -.512) = 3.02 after rotation. For component 2 the SSL’s are 2.62 and 2.91. After rotation component 1 accounted for 3.02/7 = 43% of the total variance and 3.02 / (3.02 + 2.91) = 51% of the variance distributed between the two components. After rotation the two components together account for (3.02 + 2.91)/7 = 85% of the total variance.

****

 The SSL’s for components can be used to help decide how many components to retain. An after rotation SSL is much like an eigenvalue. A rotated component with an SSL of 1 accounts for as much of the total variance as does a single variable. One may want to retain and rotate a few more components than indicated by “eigenvalue 1 or more” criterion. Inspection of the retained components’ SSL’s after rotation should tell you whether or not they should be retained. Sometimes a component with an eigenvalue > 1 will have a postrotation SSL < 1, in which case you may wish to drop it and ask for a smaller number of retained components.

 You also should look at the postrotation loadings to decide how well each retained component is defined. If only one variable loads heavily on a component, that component is not well defined. If only two variables load heavily on a component, the component may be reliable if those two variables are highly correlated with one another but not with the other variables.

**Naming Components**

 Now let us look at the rotated loadings again and try to name the two components. Component 1 has heavy loadings (>.4) on TASTE, AROMA, and COLOR and a moderate negative loading on REPUTATION. I’d call this component AESTHETIC QUALITY. Component 2 has heavy loadings on large SIZE, high ALCOHOL content, and low COST and a moderate negative loading on REPUTATION. I’d call this component CHEAP DRUNK.

**Communalities**

 Let us also look at the **SSL for each variable across factors**. Such a SSL is called a **communality**. This is the amount of the variable’s variance that is accounted for by the components (since the loadings are correlations between variables and components and the components are orthogonal, a variable’s communality represents the *R2* of the variable predicted from the components). For our beer data the communalities are COST, .84; SIZE, .90; ALCOHOL, .89; REPUTAT, .55; COLOR, .91; AROMA, .92; and TASTE, .92.

****

**Orthogonal Versus Oblique Rotations**

 The rotation I used on these data is the **VARIMAX rotation**. It is the most commonly used rotation. Its goal is to minimize the complexity of the components by making the large loadings larger and the small loadings smaller within each component. There are other rotational methods. **QUARTIMAX** rotation makes large loadings larger and small loadings smaller within each variable. **EQUAMAX** rotation is a compromise that attempts to simplify both components and variables. These are all orthogonal rotations, that is, the axes remain perpendicular, so the components are not correlated with one another.

 It is also possible to employ **oblique rotational methods**. These methods do not produce orthogonal components. Suppose you have done an orthogonal rotation and you obtain a rotated loadings plot that looks like this:

 The cluster of points midway between axes in the upper left quadrant indicates that a third component is present. The two clusters in the upper right quadrant indicate that the data would be better fit with axes that are not orthogonal. Axes drawn through those two clusters would not be perpendicular to one another. We shall return to the topic of oblique rotation later.

[Return to Multivariate Analysis with SPSS](http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-MV.htm)

[Continue on to Factor Analysis with SPSS](http://core.ecu.edu/psyc/wuenschk/MV/FA/FA-SPSS.pdf)

Copyright 2012, [Karl L. Wuensch](http://core.ecu.edu/psyc/wuenschk/wuensch.html) - All rights reserved.

1. © Copyright 2012, [Karl L. Wuensch](http://core.ecu.edu/psyc/wuenschk/wuensch.html) - All rights reserved. [↑](#footnote-ref-1)