Chapter
12

Factor Analysis

Herb Stenson and Leland Wilkinson

FACTOR provides principal components analysis and common factor analysis (maximum likelihood and iterated principal axis). SYSTAT has options to rotate, sort, plot, and save factor loadings. With the principal components method, you can also save the scores and coefficients. Orthogonal methods of rotation include varimax, equamax, quartimax, and orthomax. A direct oblimin method is also available for oblique rotation. Users can explore other rotations by interactively rotating a 3-D Quick Graph plot of the factor loadings. Various inferential statistics (for example, confidence intervals, standard errors, and chi-square tests) are provided, depending on the nature of the analysis that is run.

Resampling procedures are available in this feature.

Statistical Background

Principal components (PCA) and common factor (MLA for maximum likelihood and IPA for iterated principal axis) analyses are methods of decomposing a correlation or covariance matrix. Although principal components and common factor analyses are based on different mathematical models, they can be used on the same data and both usually produce similar results. Factor analysis is often used in exploratory data analysis to:

- Study the correlations of a large number of variables by grouping the variables in "factors" so that variables within each factor are more highly correlated with variables in that factor than with variables in other factors.
- Interpret each factor according to the meaning of the variables.

- Summarize many variables by a few factors. The scores from the factors can be used as input data for *t* tests, regression, ANOVA, discriminant analysis, and so on.

Often the users of factor analysis are overwhelmed by the gap between theory and practice. In this chapter, we try to offer practical hints. It is important to realize that you may need to make several passes through the procedure, changing options each time, until the results give you the necessary information for your problem.

If you understand the component model, you are on the way toward understanding the factor model, so let us begin with the former.

A Principal Component

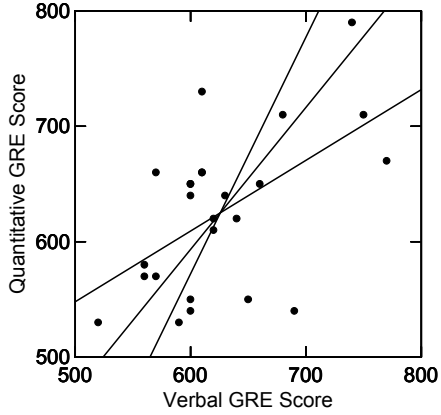
What is a principal component? The simplest way to see is through real data. The following data consist of Graduate Record Examination verbal and quantitative scores. These scores are from 25 applicants to a graduate psychology department.

VERBAL	QUANTITATIVE
590	530
620	620
640	620
650	550
620	610
610	660
560	570
610	730
600	650
740	790
560	580
680	710
600	540
520	530
660	650
750	710
630	640
570	660
600	650
570	570
600	550
690	540

VERBAL	QUANTITATIVE
770	670
610	660
600	640

Now, we could decide to try linear regression to predict verbal scores from quantitative. Or, we could decide to predict quantitative from verbal by the same method. The data does not suggest which is a dependent variable; either will do. What if we are not interested in predicting either one separately but instead want to know how both variables hang together jointly? This is what a principal component does. Karl Pearson, who developed principal component analysis in 1901, described a component as a "line of closest fit to systems of points in space." In short, the regression line indicates best *prediction*, and the component line indicates best *association*.

The following figure shows the regression and component lines for our *GRE* data. The regression of *y* on *x* is the line with the smallest slope (flatter than diagonal). The regression of *x* on *y* is the line with the largest slope (steeper than diagonal). The component line is between the other two. Interestingly, when most people are asked to draw a line relating two variables in a scatterplot, they tend to approximate the component line. It takes a lot of explaining to get them to realize that this is not the best line for predicting the vertical axis variable (*y*) or the horizontal axis variable (*x*).



Notice that the slope of the component line is approximately 1, which means that the two variables are weighted almost equally (assuming the axis scales are the same). We could make a new variable called *GRE* that is the sum of the two tests:

GRE = VERBAL + QUANTITATIVE

This new variable could summarize, albeit crudely, the information in the other two. If the points clustered almost perfectly around the component line, then the new component variable could summarize almost perfectly both variables.

Multiple Principal Components

The goal of principal components analysis is to summarize a multivariate data set as accurately as possible using a few components. So far, we have seen only one component. It is possible, however, to draw a second component perpendicular to the first. The first component will summarize as much of the joint variation as possible. The second will summarize what is left. If we do this with the *GRE* data, of course, we will have as many components as original variables—not much of a saving. We usually seek fewer components than variables, so that the variation left over is negligible.

Component Coefficients

In the above equation for computing the first principal component on our test data, we made both coefficients equal. In fact, when you run the sample covariance matrix using factor analysis in SYSTAT, the coefficients are as follows:

GRE = 0.008 * *VERBAL* + 0.01 * *QUANTITATIVE*

They are indeed nearly equal. Their magnitude is considerably less than 1 because principal components are usually scaled to conserve variance. That is, once you compute the components with these coefficients, the total variance on the components is the same as the total variance on the original variables.

Component Loadings

Most researchers want to know the relation between the original variables and the components. Some components may be nearly identical to an original variable; in other words, their coefficients may be nearly 0 for all variables except one. Other components may be a more even amalgam of several original variables.

Component loadings are the covariances of the original variables with the components. In our example, these loadings are 51.085 for *VERBAL* and 62.880 for *QUANTITATIVE*. You may have noticed that these are proportional to the coefficients;

they are simply scaled differently. If you square each of these loadings and add them up separately for each component, you will have the variance accounted for by each component.

Correlations or Covariances

Most researchers prefer to analyze the correlation rather than covariance structure among their variables. Sample correlations are simply covariances of sample standardized variables. Thus, if your variables are measured on very different scales or if you feel the standard deviations of your variables are not theoretically significant, you will want to work with correlations instead of covariances. In our test example, working with correlations yields loadings of 0.879 for each variable instead of 51.085 and 62.880. When you factor the correlation instead of the covariance matrix, then the loadings are the correlations of each component with each original variable.

For our test data, loadings of 0.879 mean that if you created a *GRE* component by standardizing *VERBAL* and *QUANTITATIVE* and adding them together weighted by the coefficients, you would find the correlation between these component scores and the original *VERBAL* scores to be 0.879. The same would be true for *QUANTITATIVE*.

Signs of Component Loadings

The signs of loadings within components are arbitrary. If a component (or factor) has more negative than positive loadings, you may change minus signs to plus and plus to minus. SYSTAT does this automatically for components that have more negative than positive loadings, and thus will occasionally produce components or factors that have different signs from those in other computer programs. This occasionally confuses users. In mathematical terms, $Ax = \lambda x$ and $-Ax = -\lambda x$ are equivalent.

Factor Analysis

We have seen how principal components analysis is a method for computing new variables that summarize variation in a space parsimoniously. For our test variables, the equation for computing the first component was:

GRE = 0.008 * *VERBAL* + 0.01 * *QUANTITATIVE*

This component equation is linear, of the form:

Component = Linear combination of {Observed variables}

Factor analysts turn this equation around:

Observed variable = Linear combination of {Factors} + Error

This model was presented by Spearman near the turn of the century in the context of a single intelligence factor and extended to multiple mental measurement factors by Thurstone several decades later. Notice that the factor model makes observed variables a function of unobserved factors. Even though this looks like a linear regression model, none of the graphical and analytical techniques used for regression can be applied to the factor model because there is no unique, observable set of factor scores or residuals to examine.

Factor analysts are less interested in prediction than in decomposing a covariance matrix. This is why the fundamental equation of factor analysis is not the above linear model, but rather its **quadratic form**:

Observed covariances = Factor covariances + Error covariances

The covariances in this equation are usually expressed in matrix form, so that the model decomposes an observed covariance matrix into a hypothetical factor covariance matrix plus a hypothetical error covariance matrix. The diagonals of these two hypothetical matrices are known, respectively, as **communalities** and **specificities**.

In ordinary language, then, the factor model expresses variation within and relations among observed variables as partly common variation among factors and partly specific variation among random errors.

Estimating Factors

Factor analysis involves several steps:

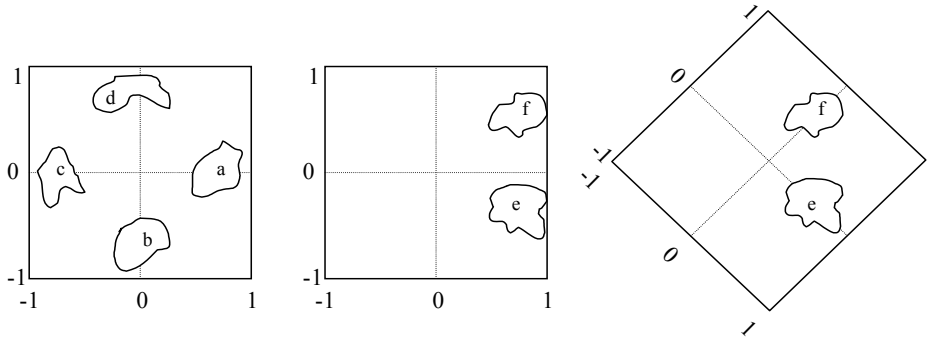
- First, the correlation or covariance matrix is computed from the usual cases-by-variables data file or it is input as a matrix.
- Second, the factor loadings are estimated. This is called **initial factor extraction**. Extraction methods are described in this section.
- Third, the factors are rotated to make the loadings more interpretable—that is, rotation methods make the loadings for each factor either large or small, not in-between. These methods are described in the next section.

Factors must be estimated iteratively in a computer. There are several methods available. The most popular approach, available in SYSTAT, is to modify the diagonal of the observed covariance matrix and calculate factors the same way components are computed. This procedure is repeated until the communalities reproduced by the factor covariances are indistinguishable from the diagonal of the modified matrix.

Rotation

Usually the initial factor extraction does not give interpretable factors. One of the purposes of rotation is to obtain factors that can be named and interpreted. That is, if you can make the large loadings larger than before and the smaller loadings smaller, then each variable is associated with a minimal number of factors. Hopefully, the variables that load strongly together on a particular factor will have a clear meaning with respect to the subject area at hand.

It helps to study plots of loadings for one factor against those for another. Ideally, you want to see clusters of loadings at extreme values for each factor: like what A and C are for factor 1, and B and D are for factor 2 in the left plot, and not like E and F in the middle plot.



In the middle plot, the loadings in groups E and F are sizeable for *both* factors 1 and 2. However, if you lift the plot axes away from E and F, rotating them 45 degrees, and then set them down as on the right, you achieve the desired effect. Sounds easy for two factors. For three factors, imagine that the loadings are balls floating in a room and that you rotate the floor and walls so that each loading is as close to the floor or a wall as it can be. This concept generalizes to more dimensions.

Researchers let the computer do the rotation automatically. There are many criteria for achieving a *simple structure* among component loadings, although Thurstone's are most widely cited. For *p* variables and *m* components:

- Each component should have at least *m* near-zero loadings.
- Few components should have nonzero loadings on the same variable.

SYSTAT provides five methods of rotating loadings: varimax, equamax, quartimax, orthomax, and oblimin.

Principal Components versus Factor Analysis

SYSTAT can perform both principal components and common factor analysis. Some view principal components analysis as a method of factor analysis, although there is a theoretical distinction between the two. Principal components are weighted linear composites of observed variables. Common factors are unobserved variables that are hypothesized to account for the intercorrelations among observed variables.

One significant practical difference is that common factor scores are indeterminate, whereas principal component scores are not. There are no sufficient estimators of scores for subjects on common factors (rotated or unrotated, maximum likelihood, or otherwise). Some computer models provide "regression" estimates of factor scores,

but these are not estimates in the usual statistical sense. This problem arises not because factors can be arbitrarily rotated (so can principal components), but because the common factor model is based on more unobserved parameters than observed data points, an unusual circumstance in statistics.

In recent years, "maximum likelihood" factor analysis algorithms have been devised to estimate common factors. The implementation of these algorithms in popular computer packages has led some users to believe that the factor indeterminacy problem does not exist for "maximum likelihood" factor estimates. It does.

Mathematicians and psychometricians have known about the factor indeterminacy problem for decades. For a historical review of the issues, see Steiger (1979); for a general review, see Rozeboom (1982). For further information refer Harman (1976), Mulaik (1972), Gnanadesikan (1977), or Mardia, Kent, and Bibby (1979), Afifi, May, and Clark (2004), Clarkson and Jennrich (1988), or Dixon (1992).

Because of the indeterminacy problem, SYSTAT computes subjects' scores only for the principal components model where subjects' scores are a simple linear transformation of scores on the factored variables. SYSTAT does not save scores from a common factor model.

Applications and Caveats

While there is not room here to discuss more statistical issues, you should realize that there are several myths about factors versus components:

Myth. The factor model allows hypothesis testing; the component model does not.
Fact. Morrison (2004) and others present a full range of formal statistical tests for components.

Myth. Factor loadings are real; principal component loadings are approximations.
Fact. This statement is too ambiguous to have any meaning. It is easy to define things so that factors are approximations of components.

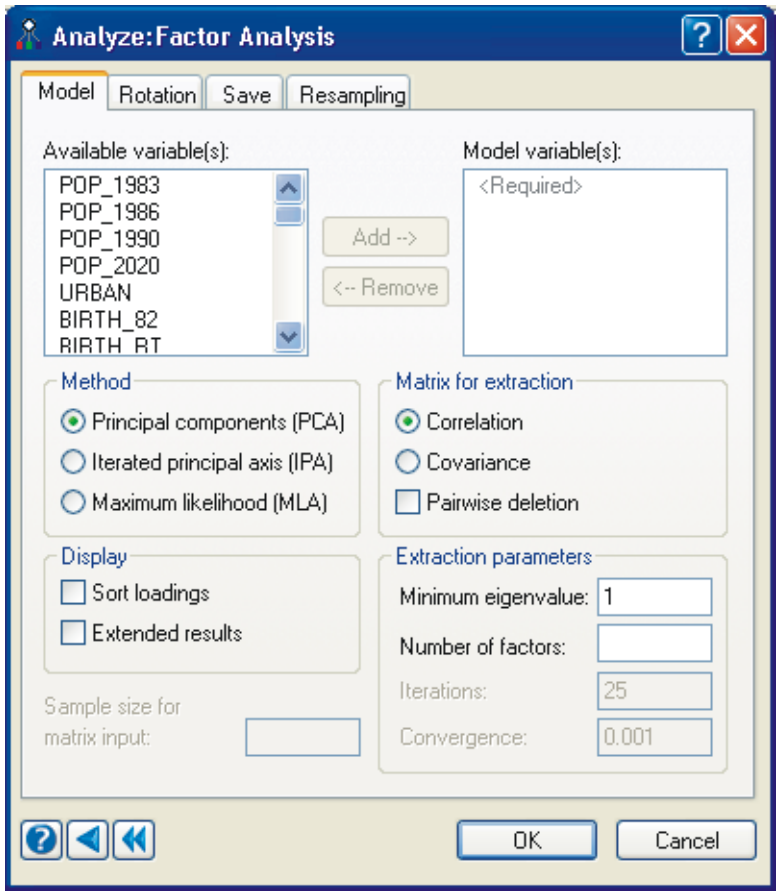
Myth. Factor analysis is more likely to uncover lawful structure in your data; principal components are more contaminated by error.
Fact. Again, this statement is ambiguous. With further definition, it can be shown to be true for some data, false for other. It is true that, in general, factor solutions will have lower dimensionality than corresponding component solutions. This can be an advantage when searching for simple structure among noisy variables, as long as you compare the result to a principal components solution to avoid being fooled by the sort of degeneracies illustrated above.

Factor Analysis in SYSTAT

Factor Analysis Dialog Box

For factor analysis, from the menus choose:

Analyze
  Factor Analysis…



The following options are available:

Model variables. Variables used to create factors.

Method. SYSTAT offers three estimation methods:

- Principal components analysis (PCA) is the default method of analysis.
- Iterated principal axis (IPA) provides an iterative method to extract common factors by starting with the principal components solution and iteratively solving for communalities.
- Maximum likelihood analysis (MLA) iteratively finds communalities and common factors.

Display. You can sort factor loadings by size or display extended results. Selecting Extended results displays all possible factor output.

Sample size for matrix input. If your data are in the form of a correlation or covariance matrix, you must specify the sample size on which the input matrix is based so that inferential statistics (available with extended results) can be computed.

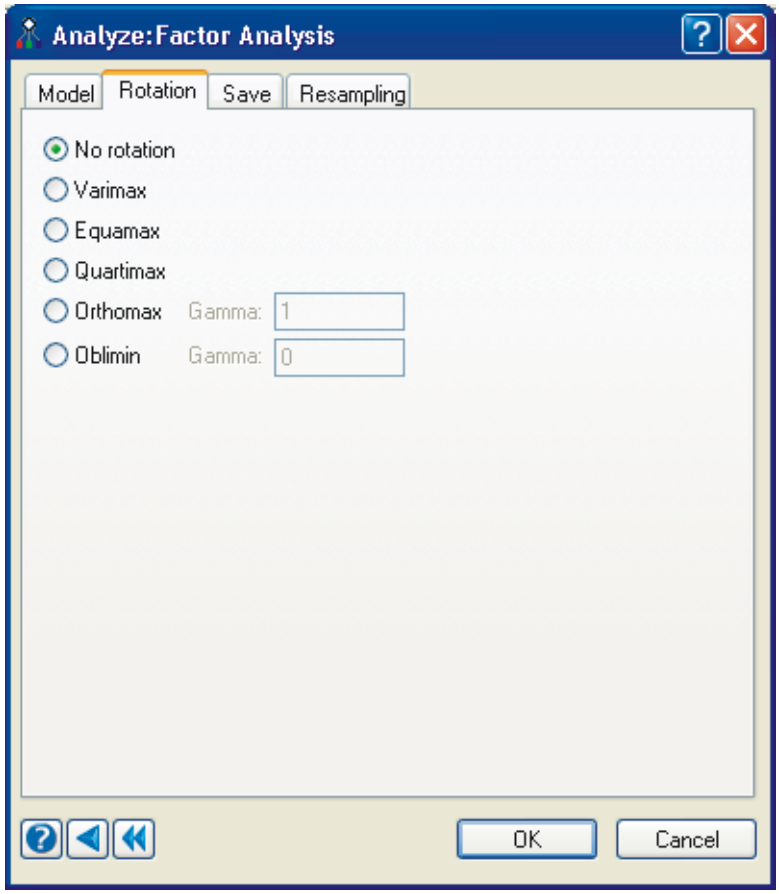
Matrix for extraction. You can factor a correlation matrix or a covariance matrix. Most frequently, the correlation matrix is used. You can also delete missing cases pairwise instead of listwise. Listwise deletes any case with missing data for any variable in the list. Pairwise examines each pair of variables and uses all cases with both values present.

Extraction parameters. You can limit the results by specifying extraction parameters.

- **Minimum eigenvalue.** Specify the smallest eigenvalue to retain. The default is 1.0 for PCA and IPA (not available with maximum likelihood). Incidentally, if you specify 0, factor analysis ignores components with negative eigenvalues (which can occur with pairwise deletion).
- **Number of factors.** Specify the number of factors to compute. If you specify both the number of factors and the minimum eigenvalue, factor analysis uses whichever criterion results in the smaller number of components.
- **Iterations.** Specify the number of iterations SYSTAT should perform (not available for principal components). The default is 25.
- **convergence.** Specify the convergence criterion (not available for principal components). The default is 0.001.

Rotation

This tab specifies the factor rotation method.



The following methods are available:

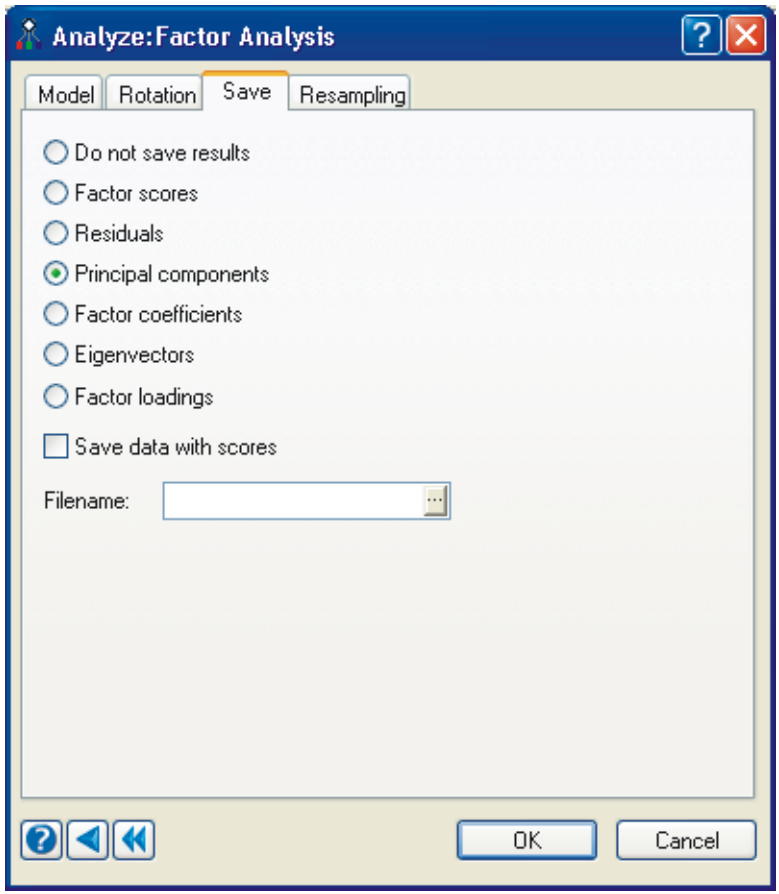
- **No rotation**. Factors are not rotated.
- **Varimax**. An orthogonal rotation method that minimizes the number of variables that have high loadings on each factor. It simplifies the interpretation of the factors.
- **Equamax**. A rotation method that is a combination of the varimax method, which simplifies the factors, and the quartimax method, which simplifies the variables.

The number of variables that load highly on a factor and the number of factors needed to explain a variable are minimized.

- **Quartimax**. A rotation method that minimizes the number of factors needed to explain each variable. It simplifies the interpretation of the observed variables.
- **Orthomax**. Specifies families of orthogonal rotations. Gamma specifies the member of the family to use. Varying Gamma changes maximization of the variances of the loadings from columns (Varimax) to rows (Quartimax).
- **Oblimin**. Specifies families of oblique (non-orthogonal) rotations. Gamma specifies the member of the family to use. For Gamma, specify 0 for moderate correlations, positive values to allow higher correlations, and negative values to restrict correlations.

Save

You can save factor analysis results for further analyses.



For the maximum likelihood and iterated principal axis methods, you can save only loadings. For the principal components method, select from these options:

- **Do not save results.** Results are not saved.
- **Factor scores.** Standardized factor scores
- **Residuals.** Residuals for each case. For a correlation matrix, the residual is the actual *z* score minus the predicted *z* score using the factor scores times the loadings to get the predicted scores. For a covariance matrix, the residuals are from

unstandardized predictions. With an orthogonal rotation, Q and PROB are also saved. Q is the sum of the squared residuals, and PROB is its probability.

- **Principal components.** Unstandardized principal components scores with mean 0 and variance equal to the eigenvalue for the factor (only for PCA without rotation).
- **Factor coefficients.** Coefficients that produce standardized scores. For a correlation matrix, multiply the coefficients by the standardized variables; for a covariance matrix, use the original variables.
- **Eigenvectors.** Eigenvectors (only for PCA without a rotation). Use to produce unstandardized scores.
- **Factor loadings.** Factor loadings.
- **Save data with scores.** Saves the selected item and all the variables in the working data file as a new data file. Use with options for scores (not loadings, coefficients, or other similar options).

If you save scores, the variables in the file are labeled *FACTOR(1)*, *FACTOR(2)*, and so on. Any observations with missing values on any of the input variables will have missing values for all scores. The scores are normalized to have zero mean and, if the correlation matrix is used, unit variance. If you use the covariance matrix and perform no rotations, SYSTAT does not standardize the component scores. The sum of their variances is the same as for the original data.

If you want to use the score coefficients to get component scores for new data, multiply the coefficients by the standardized data. SYSTAT does this when it saves scores. Another way to do cross-validation is to assign a zero weight to those cases not used in the factoring and to assign a unit weight to those cases used. The zero-weight cases are not used in the factoring, but scores are computed for them.

When Factor scores or Principal components is requested, *T2* and *PROB* are also saved. The former is the Hotelling T^2 statistic that squares the standardized distance from each case to the centroid of the factor space (that is, the sum of the squared, standardized factor scores). *PROB* is the upper-tail probability of *T2*. Use this statistic to identify outliers within the factor space. *T2* is not computed with an oblique rotation.

Using Commands

After selecting a data file with USE filename, continue with:

```
FACTOR
MODEL varlist
SAVE filename / SCORES  DATA  LOAD  COEF  VECTORS  PC  RESID
ESTIMATE / METHOD = PCA or IPA or MLA ,
           LISTWISE or PAIRWISE  N=n  CORR or COVA ,
           NUMBER=n  EIGEN=n  ITER=n  CONV=n  SORT ,
           ROTATE = VARIMAX or EQUAMAX or QUARTIMAX
                    or ORTHOMAX or OBLIMIN
           GAMMA=n SAMPLE = BOOT(m,n) JACK SIMPLE(m,n)
```

Usage Considerations

Types of data. Data for factor analysis can be a cases-by-variables data file, a correlation matrix, or a covariance matrix.

Print options. Factor analysis offers three categories of output: Short (the default), Medium, and Long. Each has specific output panels associated with it.

For Short, the default, panels are: Latent roots or eigenvalues (not MLA), initial communality estimates (not PCA), component loadings (PCA) or factor pattern (MLA, IPA), variance explained by components (PCA) or factors (MLA, IPA), percentage of total variance explained, change in uniqueness and log likelihood at each iteration (MLA only), and canonical correlations (MLA only). When a rotation is requested: rotated loadings (PCA) or pattern (MLA, IPA) matrix, variance explained by rotated components, percentage of total variance explained, and correlations among oblique components or factors (oblimin only).

By specifying Medium, you get the panels listed for Short, plus: the matrix to factor, the chi-square test that all eigenvalues are equal (PCA only), the chi-square test that the last *k* eigenvalues are equal (PCA only), and differences of original correlations or covariances minus fitted values. For covariance matrix input (not MLA or IPA): asymptotic 95% confidence limits for the eigenvalues and estimates of the population eigenvalues with standard errors.

With Long, you get the panels listed for Short and Medium, plus: latent vectors (eigenvectors) with standard errors (not MLA) and the chi-square test that the number of factors is *k* (MLA only) and factor coefficients. With an oblimin rotation: direct and indirect contribution of factors to variances and the rotated structure matrix.

Quick Graphs. Factor analysis produces a scree plot and a factor loadings plot.

Saving files. You can save factor scores, residuals, principal components, factor coefficients, eigenvectors, or factor loadings as a new data file. For the iterated principal axis and maximum likelihood methods, you can save only factor loadings. You can save only eigenvectors and principal components for unrotated solutions using the principal components method.

BY groups. Factor analysis produces separate analyses for each level of any BY variables.

Case frequencies. Factor analysis uses FREQUENCY variables to duplicate cases for rectangular data files.

Case weights. For rectangular data, you can weight cases using a WEIGHT variable.

Examples

Example 1
Principal Components

Principal components (PCA, the default method) is a good way to begin a factor analysis (and possibly the only method you may need). If one variable is a linear combination of the others, the program will not stop (MLA and IPA both require a nonsingular correlation or covariance matrix). The PCA output can also provide indications that:

- One or more variables have little relation to the others and, therefore, are not suited for factor analysis - so in your next run, you might consider omitting them.
- The final number of factors may be three or four and not double or triple this number.

To illustrate this method of factor extraction, we borrow data from Harman (1976), who borrowed them from a 1937 unpublished thesis by Mullen. This classic data set is widely used in the literature. For example, Jackson (2003) reports loadings for the PCA, MLA, and IPA methods. The data are measurements recorded for 305 youth aged seven to seventeen: height, arm span, length of forearm, length of lower leg, weight,

bitrochanteric diameter (the upper thigh), girth, and width. Because the units of these measurements differ, we analyze a correlation matrix:

	Height	Arm_Span	Forearm	Lowerleg	Weight	Bitro	Girth	Width
Height	1.000							
Arm_Span	0.846	1.000						
Forearm	0.805	0.881	1.000					
Lowerleg	0.859	0.826	0.801	1.000				
Weight	0.473	0.376	0.380	0.436	1.000			
Bitro	0.398	0.326	0.319	0.329	0.762	1.000		
Girth	0.301	0.277	0.237	0.327	0.730	0.583	1.000	
Width	0.382	0.415	0.345	0.365	0.629	0.577	0.539	1.000

The correlation matrix is stored in the *YOUTH* file. SYSTAT knows that the file contains a correlation matrix, so no special instructions are needed to read the matrix.

The input is:

```
FACTOR
   USE YOUTH
   MODEL HEIGHT..WIDTH
   ESTIMATE / METHOD=PCA  N=305  SORT  ROTATE=VARIMAX
```

Notice the shortcut notation (..) for listing consecutive variables in a file.
The output is:

```
Latent Roots (Eigenvalues)
  1        2        3        4        5
 ---------------------------------------------
  4.6729   1.7710   0.4810   0.4214   0.2332

  6        7        8
 -------------------------
  0.1867   0.1373   0.0965

  Component Loadings

            |      1         2
 ----------+-----------------
  HEIGHT    | 0.8594    0.3723
  ARM_SPAN  | 0.8416    0.4410
  LOWERLEG  | 0.8396    0.3953
  FOREARM   | 0.8131    0.4586
  WEIGHT    | 0.7580   -0.5247
  BITRO     | 0.6742   -0.5333
  WIDTH     | 0.6706   -0.4185
  GIRTH     | 0.6172   -0.5801
```

Variance Explained by Components

1	2
4.6729	1.7710

Percent of Total Variance Explained

1	2
58.4110	22.1373

Rotated Loading Matrix (VARIMAX, Gamma = 1.000000)

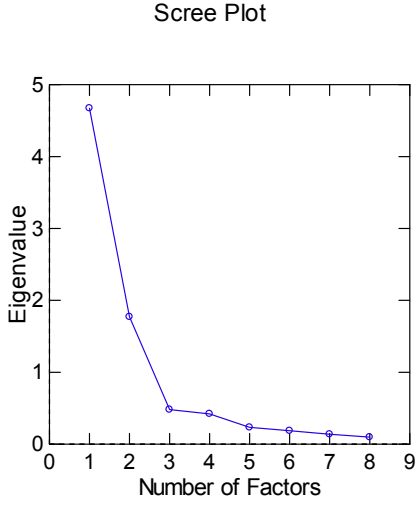
	1	2
ARM_SPAN	0.9298	0.1955
FOREARM	0.9191	0.1638
HEIGHT	0.8998	0.2599
LOWERLEG	0.8992	0.2295
WEIGHT	0.2507	0.8871
BITRO	0.1806	0.8404
GIRTH	0.1068	0.8403
WIDTH	0.2509	0.7496

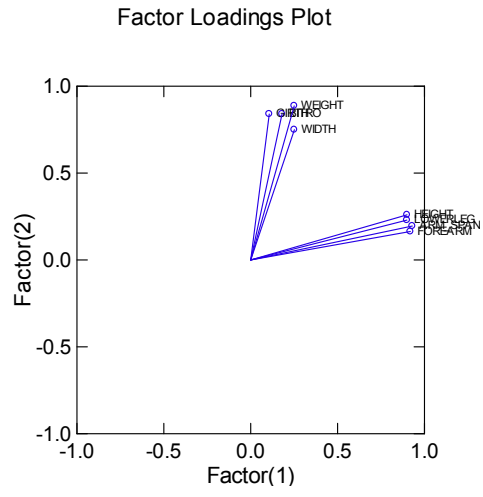
"Variance" Explained by Rotated Components

1	2
3.4973	2.9465

Percent of Total Variance Explained

1	2
43.7165	36.8318





Notice that we did not specify how many factors we wanted. For PCA, the assumption

is to compute as many factors as there are eigenvalues greater than 1.0—so, in this run, you study results for two factors. After examining the output, you may want to specify a minimum eigenvalue or, very rarely, a lower limit.

Unrotated loadings (and orthogonally rotated loadings) are correlations of the variables with the principal components (factors). They are also the eigenvectors of the correlation matrix multiplied by the square roots of the corresponding eigenvalues. Usually these loadings are not useful for interpreting the factors. For some industrial applications, researchers prefer to examine the eigenvectors alone.

The *Variance explained* for each component is the eigenvalue for the factor. The first factor accounts for 58.4% of the variance; the second, 22.1%. The *Total Variance* is the sum of the diagonal elements of the correlation (or covariance) matrix. By summing the *Percent of Total Variance Explained* for the two factors ($58.411 + 22.137 = 80.548$), you can say that more than 80% of the variance of all eight variables is explained by the first two factors.

In the *Rotated Loading Matrix*, the rows of the display have been sorted, placing the loadings > 0.5 for factor 1 first, and so on. These are the coefficients of the factors after rotation, so notice that large values for the unrotated loadings are larger here and the small values are smaller. The sum of squares of these coefficients (for each factor or column) are printed below under the heading *Variance Explained by Rotated Components*. Together, the two rotated factors explain more than 80% of the variance. Factor analysis offers five types of rotation. Here, by default, the orthogonal varimax method is used.

To interpret each factor, look for variables with high loadings. The four variables that load highly on factor 1 can be said to measure "lankiness"; while the four that load highly on factor 2, "stockiness." Other data sets may include variables that do not load highly on any specific factor.

In the factor scree plot, the eigenvalues are plotted against their order (or associated component). Use this display to identify large values that separate well from smaller eigenvalues. This can help to identify a useful number of factors to retain. Scree is the rubble at the bottom of a cliff; the large retained roots are the cliff, and the deleted ones are the rubble.

The points in the factor loadings plot are variables, and the coordinates are the rotated loadings. Look for clusters of loadings at the extremes of the factors. The four variables at the right of the plot load highly on factor 1 and all reflect length. The variables at the top of the plot load highly on factor 2 and reflect width.

Example 2
Maximum Likelihood

This example uses maximum likelihood for initial factor extraction and 2 as the number of factors. Other options remain as in the principal components example.

The input is:

```
FACTOR
   USE YOUTH
   MODEL HEIGHT..WIDTH
   ESTIMATE / METHOD=MLA  N=305  NUMBER=2  SORT  ROTATE=VARIMAX
```

The output is:

Initial Communality Estimates

```
1        2        3        4        5        6        7        8
-------------------------------------------------------------------------
0.8162   0.8493   0.8006   0.7884   0.7488   0.6041   0.5622   0.4778

Iterative Maximum Likelihood Factor Analysis: Convergence = 0.0010.
```

Iterations History

Iteration Number	Maximum Change in SQRT (uniqueness)	Negative log of Likelihood
1	0.7226	0.3841
2	0.2438	0.2733
3	0.0512	0.2537
4	0.0104	0.2532
5	0.0005	0.2532

Canonical Correlations

1	2
0.9823	0.9489

Factor Pattern

	1	2	Communality Estimates	Specific Variances
HEIGHT	0.8797	0.2375	0.8302	0.1698
ARM_SPAN	0.8735	0.3604	0.8929	0.1071
LOWERLEG	0.8551	0.2633	0.8006	0.1994
FOREARM	0.8458	0.3442	0.8338	0.1662
WEIGHT	0.7048	-0.6436	0.9109	0.0891
BITRO	0.5887	-0.5383	0.6363	0.3637
WIDTH	0.5743	-0.3653	0.4633	0.5367
GIRTH	0.5265	-0.5536	0.5837	0.4163

Variance Explained by Factors

1	2
4.4337	1.5179

Percent of Total Variance Explained

1	2
55.4218	18.9742

Rotated Pattern Matrix (VARIMAX, Gamma = 1.000000)

	1	2
ARM_SPAN	0.9262	0.1873
FOREARM	0.8942	0.1853
HEIGHT	0.8628	0.2928
LOWERLEG	0.8569	0.2576
WEIGHT	0.2268	0.9271
BITRO	0.1891	0.7750
GIRTH	0.1289	0.7530
WIDTH	0.2734	0.6233

"Variance" Explained by Rotated Factors

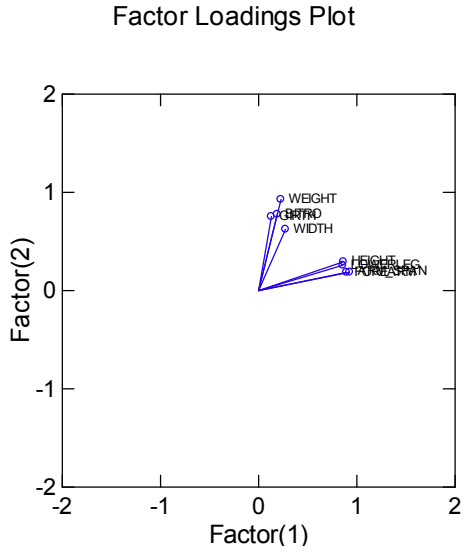
1	2
3.3146	2.6370

Percent of Total Variance Explained

1	2
41.4331	32.9628

Percent of Common Variance Explained

1	2
55.6927	44.3073



The first panel of output contains the communality estimates. The communality of a variable is its theoretical squared multiple correlation with the factors extracted. For MLA (and IPA), the assumption for the initial communalities is the observed squared multiple correlation with all the other variables.

The canonical correlations are the largest multiple correlations for successive orthogonal linear combinations of factors with successive orthogonal linear combinations of variables. These values are comfortably high. If, for other data, some of the factors have values that are much lower, you might want to request fewer factors.

The loadings and amount of variance explained are similar to those found in the principal components example. In addition, maximum likelihood reports the percentage of common variance explained. Common variance is the sum of the communalities. If **A** is the unrotated MLA factor pattern matrix, common variance is the trace of $\mathbf{A}'\mathbf{A}$.

Number of Factors

In this example, we specified two factors to extract. If you were to omit this specification and rerun the example, SYSTAT adds this report to the output

```
The Maximum Number of Factors for your Data is 4.
```

SYSTAT will also report this message if you request more than four factors for these data. This result is due to a theorem by Lederman and indicates that the degrees of freedom allow estimates of loadings and communalities for only four factors.

If we set the print length to long, SYSTAT reports:

```
Chi-square Test that the Number of Factors is 4

Chi-square : 4.3187
df         : 2.0000
p-value    : 0.1154
```

The results of this chi-square test indicate that you do not reject the hypothesis that there are four factors (*p-value* > 0.05). Technically, the hypothesis is that “no more than

four factors are required." This, of course, does not negate 2 as the right number. For the *YOUTH* data, here are rotated loadings for four factors:

Rotated Pattern Matrix (VARIMAX, Gamma = 1.000000)

	1	2	3	4
ARM_SPAN	0.9372	0.1984	-0.2831	0.0465
LOWERLEG	0.8860	0.2142	0.1878	0.1356
HEIGHT	0.8776	0.2819	0.1134	-0.0077
FOREARM	0.8732	0.1957	-0.0851	-0.0065
WEIGHT	0.2414	0.8830	0.1077	0.1080
BITRO	0.1823	0.8233	0.0163	-0.0784
GIRTH	0.1133	0.7315	-0.0048	0.5219
WIDTH	0.2597	0.6459	-0.1400	0.0819

The loadings for the last two factors do not make sense. Possibly, the fourth factor has one variable, GIRTH, but it still has a healthier loading on factor 2. This test is based on an assumption of multivariate normality (as is MLA itself). If not true, then the test is invalid.

Example 3
Iterated Principal Axis

This example continues with the *YOUTH* data described in the principal components example, this time using the IPA (iterated principal axis) method to extract factors.

The input is:

```
FACTOR
   USE YOUTH
   MODEL HEIGHT..WIDTH
   ESTIMATE / METHOD=IPA  SORT  ROTATE=VARIMAX
```

The output is:

Initial Communality Estimates

1	2	3	4	5	6	7	8
0.8162	0.8493	0.8006	0.7884	0.7488	0.6041	0.5622	0.4778

```
Iterative Maximum Likelihood Factor Analysis: Convergence = 0.0010.
```

Iterations History

Iteration Number	Maximum Change in SQRT (uniqueness)	Negative log of Likelihood
1	0.7226	0.3841
2	0.2438	0.2733
3	0.0512	0.2537
4	0.0104	0.2532
5	0.0005	0.2532

Canonical Correlations

1	2
0.9823	0.9489

Factor Pattern

	1	2	Communality Estimates	Specific Variances
HEIGHT	0.8797	0.2375	0.8302	0.1698
ARM_SPAN	0.8735	0.3604	0.8929	0.1071
LOWERLEG	0.8551	0.2633	0.8006	0.1994
FOREARM	0.8458	0.3442	0.8338	0.1662
WEIGHT	0.7048	-0.6436	0.9109	0.0891
BITRO	0.5887	-0.5383	0.6363	0.3637
WIDTH	0.5743	-0.3653	0.4633	0.5367
GIRTH	0.5265	-0.5536	0.5837	0.4163

Variance Explained by Factors

1	2
4.4337	1.5179

Percent of Total Variance Explained

1	2
55.4218	18.9742

Rotated Pattern Matrix (VARIMAX, Gamma = 1.000000)

	1	2
ARM_SPAN	0.9262	0.1873
FOREARM	0.8942	0.1853
HEIGHT	0.8628	0.2928
LOWERLEG	0.8569	0.2576
WEIGHT	0.2268	0.9271
BITRO	0.1891	0.7750
GIRTH	0.1289	0.7530
WIDTH	0.2734	0.6233

"Variance" Explained by Rotated Factors

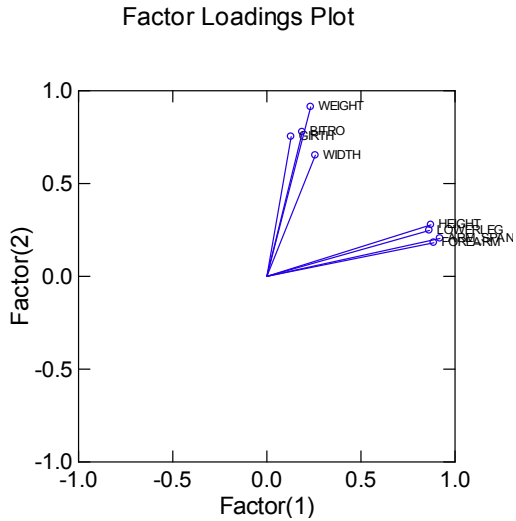
1	2
3.3146	2.6370

Percent of Total Variance Explained

1	2
41.4331	32.9628

Percent of Common Variance Explained

1	2
55.6927	44.3073



Before the first iteration, the communality of a variable is its multiple correlation squared with the remaining variables. At each iteration, communalities are estimated from the loadings matrix, **A**, by finding the trace of **A'A**, where the number of columns in **A** is the number of factors. Iterations continue until the largest change in any communality is less than that specified with *Convergence*. Replacing the diagonal of the correlation (or covariance) matrix with these final communality estimates and computing the eigenvalues yields the latent roots in the next panel.

Example 4
Rotation

Let us compare the unrotated and orthogonally rotated loadings from the principal components example with those from an oblique rotation.

The input is:

```
FACTOR
   USE YOUTH
   PLENGTH LONG
   MODEL HEIGHT..WIDTH
   ESTIMATE / METHOD=PCA  N=305  SORT

   MODEL HEIGHT..WIDTH
   ESTIMATE / METHOD=PCA  N=305  SORT  ROTATE=VARIMAX

   MODEL HEIGHT..WIDTH
   ESTIMATE / METHOD=PCA  N=305  SORT  ROTATE=OBLIMIN
```

We focus on the output directly related to the rotations.

The output is:

```
Component Loadings
           |      1         2
 ----------+-----------------
  HEIGHT   | 0.8594    0.3723
  ARM_SPAN | 0.8416    0.4410
  LOWERLEG | 0.8396    0.3953
  FOREARM  | 0.8131    0.4586
  WEIGHT   | 0.7580   -0.5247
  BITRO    | 0.6742   -0.5333
  WIDTH    | 0.6706   -0.4185
  GIRTH    | 0.6172   -0.5801

  Variance Explained by Components

  1        2
 ----------------
  4.6729   1.7710

  Percent of Total Variance Explained

  1         2
 ------------------
  58.4110   22.1373
  Rotated Loading Matrix (VARIMAX, Gamma = 1.000000)

           |      1        2
 ----------+-----------------
  ARM_SPAN | 0.9298   0.1955
  FOREARM  | 0.9191   0.1638
  HEIGHT   | 0.8998   0.2599
  LOWERLEG | 0.8992   0.2295
  WEIGHT   | 0.2507   0.8871
  BITRO    | 0.1806   0.8404
  GIRTH    | 0.1068   0.8403
  WIDTH    | 0.2509   0.7496

  "Variance" Explained by Rotated Components

  1        2
 ----------------
  3.4973   2.9465
```

Percent of Total Variance Explained

1	2
43.7165	36.8318

Rotated Pattern Matrix (OBLIMIN, Gamma = 0.000000)

	1	2
ARM_SPAN	0.9572	-0.0166
FOREARM	0.9533	-0.0482
LOWERLEG	0.9157	0.0276
HEIGHT	0.9090	0.0604
WEIGHT	0.0537	0.8975
GIRTH	-0.0904	0.8821
BITRO	-0.0107	0.8642
WIDTH	0.0876	0.7487

"Variance" Explained by Rotated Components

1	2
3.5273	2.9166

Percent of Total Variance Explained

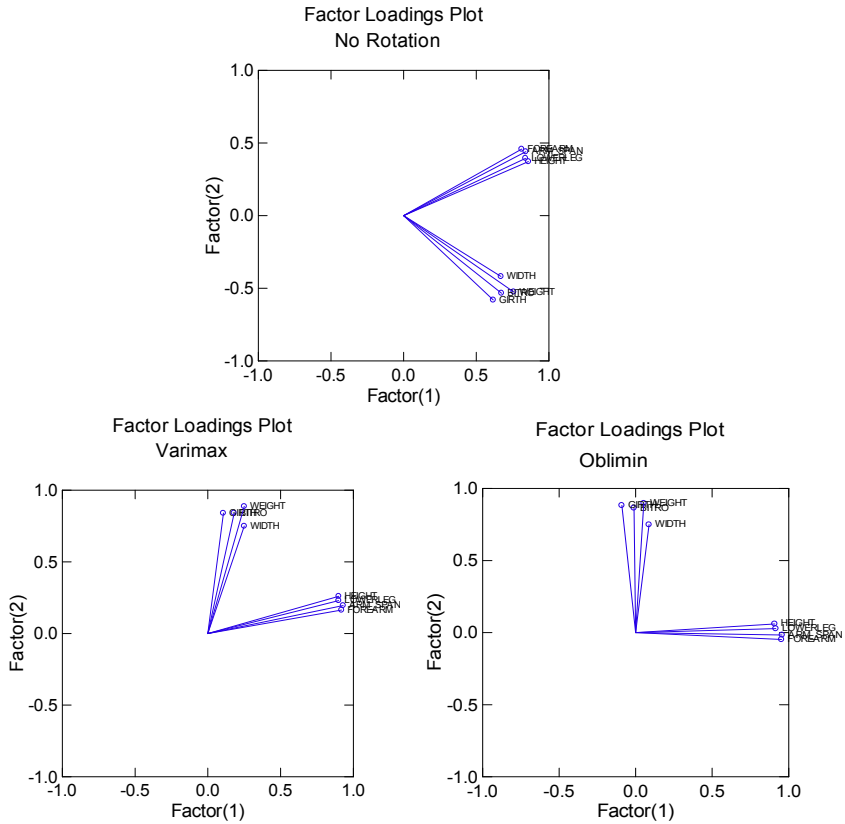
1	2
44.0913	36.4569

Direct and Indirect Contributions of Factors to Variance

	1	2
1	3.5087	
2	0.0186	2.8979

Rotated Structure Matrix

	1	2
ARM_SPAN	0.9500	0.3962
FOREARM	0.9325	0.3629
LOWERLEG	0.9277	0.4225
HEIGHT	0.9350	0.4523
WEIGHT	0.4407	0.9206
GIRTH	0.2900	0.8431
BITRO	0.3620	0.8596
WIDTH	0.4104	0.7865



The values in *Direct and Indirect Contributions of Factors to Variance* are useful for determining if a part of a factor's contribution to *"Variance" Explained* is due to its correlation with another factor. Notice that

3.5087 + 0.0186 = 3.5273

is the *"Variance" Explained* for factor 1, and

2.8979 + 0.0186 = 2.9165

is the *"Variance" Explained* for factor 2.

Think of the values in the *Rotated Structure Matrix* as correlations of the variable with the factors. Here we see that the first four variables are highly correlated with the first factor. The remaining variables are highly correlated with the second factor.

The factor loading plots illustrate the effects of the rotation methods. While the unrotated factor loadings form two distinct clusters, they both have strong positive loadings for factor 1. The "lanky" variables have moderate positive loadings on factor 2 while the "stocky" variables have negative loadings on factor 2. With the varimax rotation, the "lanky" variables load highly on factor 1 with small loadings on factor 2; the "stocky" variables load highly on factor 2. The oblimin rotation does a much better job of centering each cluster at 0 on its minor factor.

Example 5
Factor Analysis Using a Covariance Matrix

Jackson (1991) describes a project in which the maximum thrust of ballistic missiles was measured. For a specific measure called "total impulse," it is necessary to calculate the area under a curve. Originally, a planimeter was used to obtain the area, and later an electronic device performed the integration directly but unreliably in its early usage. As data, two strain gauges were attached to each of 40 Nike rockets, and both types of measurements were recorded in parallel (making four measurements per rocket). The covariance matrix of the measures is stored in the *MISSLES* file.

In this example, we illustrate features associated with covariance matrix input (asymptotic 95% confidence limits for the eigenvalues, estimates of the population eigenvalues with standard errors, and latent vectors (eigenvectors or characteristic vectors) with standard errors).

The input is:

```
FACTOR
   USE MISSLES
   MODEL INTEGRA1 PLANMTR1 INTEGRA2 PLANMTR2
   PLENGTH LONG
   ESTIMATE / METHOD=PCA  COVA  N=40
```

The output is:

```
Latent Roots (Eigenvalues)
  1         2         3         4
 ---------------------------------------
  335.3355   48.0344   29.3305   16.4096

Empirical Upper Bound for the First Eigenvalue : 398.0000

Asymptotic 95% Confidence Limits for the Eigenvalues, N = 40
```

Upper Limits

1	2	3	4
596.9599	85.5102	52.2138	29.2122

Lower Limits

1	2	3	4
233.1534	33.3975	20.3930	11.4093

Unbiased Estimates of Population Eigenvalues

1	2	3	4
332.6990	46.9298	31.0859	18.3953

Unbiased Estimates of Standard Errors of Eigenvalues

1	2	3	4
74.9460	10.1768	5.7355	3.2528

Chi-square Test that All Eigenvalues are Equal

N : 40.0000
Chi-square : 110.6871
df : 9.0000
p-value : 0.0000

Latent Vectors (Eigenvectors)

	1	2	3	4
INTEGRA1	0.4681	0.6215	0.5716	0.2606
PLANMTR1	0.6079	0.1788	-0.7595	0.1473
INTEGRA2	0.4590	-0.1387	0.1677	-0.8614
PLANMTR2	0.4479	-0.7500	0.2615	0.4104

Standard Error for Each Eigenvector Element

	1	2	3	4
INTEGRA1	0.0532	0.1879	0.2106	0.1773
PLANMTR1	0.0412	0.2456	0.0758	0.2066
INTEGRA2	0.0342	0.1359	0.2366	0.0519
PLANMTR2	0.0561	0.1058	0.2633	0.1276

Component Loadings

	1	2	3	4
INTEGRA1	8.5727	4.3072	3.0954	1.0559
PLANMTR1	11.1325	1.2389	-4.1131	0.5965
INTEGRA2	8.4051	-0.9616	0.9084	-3.4893
PLANMTR2	8.2017	-5.1983	1.4165	1.6625

Variance Explained by Components

1	2	3	4
335.3355	48.0344	29.3305	16.4096

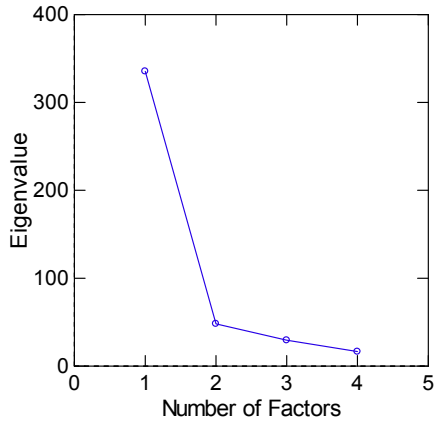
Percent of Total Variance Explained

1	2	3	4
78.1467	11.1940	6.8352	3.8241

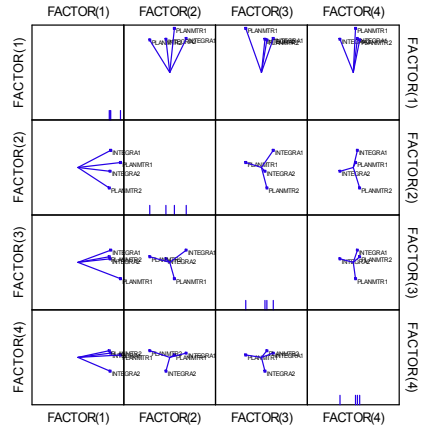
Differences: Original Minus Fitted Correlations or Covariances

	INTEGRA1	PLANMTR1	INTEGRA2	PLANMTR2
INTEGRA1	0.0000			
PLANMTR1	0.0000	0.0000		
INTEGRA2	0.0000	0.0000	0.0000	
PLANMTR2	0.0000	0.0000	0.0000	0.0000

Scree Plot



Factor Loadings Plot



SYSTAT performs a test to determine if all eigenvalues are equal. The null hypothesis is that all eigenvalues are equal against an alternative hypothesis that at least one root is different. The results here indicate that you reject the null hypothesis ($p < 0.00005$). At least one of the eigenvalues differs from the others.

The size and sign of the loadings reflect how the factors and variables are related. The first factor has fairly similar loadings for all four variables. You can interpret this factor as an overall average of the area under the curve across the four measures. The second factor represents gauge differences because the signs are different for each. The third factor is primarily a comparison between the first planimeter and the first integration device. The last factor has no simple interpretation.

When there are four or more factors, the Quick Graph of the loadings is a SPLOM. The first component represents 78% of the variability of the product, so plots of

loadings for factors 2 through 4 convey little information (notice that values in the stripe displays along the diagonal concentrate around 0, while those for factor 1 fall to the right).

Example 6
Factor Analysis Using a Rectangular File

Begin this analysis from the *OURWORLD* cases-by-variables data file. Each case contains information for one of 57 countries. We will study the interrelations among a subset of 13 variables including economic measures (gross domestic product per capita and U.S. dollars spent per person on education, health, and the military), birth and death rates, population estimates for 1983, 1986, and 1990 plus predictions for 2020, and the percentages of the population who can read and who live in cities.

We request principal components extraction with an oblique rotation. As a first step, SYSTAT computes the correlation matrix. Correlations measure linear relations. However, plots of the economic measures and population values as recorded indicate a lack of linearity, so you use base 10 logarithms to transform six variables, and you use square roots to transform two others.

The input is:

```
FACTOR
   USE OURWORLD
   LET (GDP_CAP, GNP_86, POP_1983, POP_1986, POP_1990, POP_2020),
        = L10(@)
   LET (MIL,EDUC) = SQR(@)
   MODEL URBAN BIRTH_RT DEATH_RT GDP_CAP GNP_86 MIL,
         EDUC B_TO_D LITERACY POP_1983 POP_1986,
         POP_1990 POP_2020
   PLENGTH MEDIUM
   SAVE pcascore / SCORES
   ESTIMATE / METHOD=PCA SORT ROTATE=OBLIMIN
```

The output is:

```
Matrix to be Factored
            |   URBAN   BIRTH_RT   DEATH_RT   GDP_CAP    GNP_86
 -----------+---------------------------------------------------
  URBAN     |  1.0000
  BIRTH_RT  | -0.8002     1.0000
  DEATH_RT  | -0.5126     0.5110     1.0000
  GDP_CAP   |  0.7636    -0.9189    -0.4012    1.0000
  GNP_86    |  0.7747    -0.8786    -0.4518    0.9736    1.0000
  MIL       |  0.6453    -0.7547    -0.1482    0.8657    0.8514
  EDUC      |  0.6238    -0.7528    -0.2151    0.8996    0.9207
  B_TO_D    | -0.3074     0.5106    -0.4340   -0.5293   -0.4411
  LITERACY  |  0.7997    -0.9302    -0.6601    0.8337    0.8404
```

POP_1983	0.2133	-0.0836	0.0152	0.0583	0.0090
POP_1986	0.1898	-0.0523	0.0291	0.0248	-0.0215
POP_1990	0.1700	-0.0252	0.0284	-0.0015	-0.0447
POP_2020	0.0054	0.1880	0.0743	-0.2116	-0.2484

	MIL	EDUC	B_TO_D	LITERACY	POP_1983
MIL	1.0000				
EDUC	0.8869	1.0000			
B_TO_D	-0.6184	-0.5252	1.0000		
LITERACY	0.6421	0.6869	-0.2737	1.0000	
POP_1983	0.2206	-0.0062	-0.1526	-0.0050	1.0000
POP_1986	0.1942	-0.0306	-0.1358	-0.0327	0.9984
POP_1990	0.1727	-0.0513	-0.1070	-0.0534	0.9966
POP_2020	-0.0339	-0.2555	0.0617	-0.2360	0.9531

	POP_1986	POP_1990	POP_2020
POP_1986	1.0000		
POP_1990	0.9992	1.0000	
POP_2020	0.9605	0.9673	1.0000

Latent Roots (Eigenvalues)

1	2	3	4	5
6.3950	4.0165	1.6557	0.4327	0.2390

6	7	8	9	10
0.0966	0.0812	0.0403	0.0251	0.0110

11	12	13
0.0054	0.0012	0.0002

Empirical Upper Bound for the First Eigenvalue : 7.4817

Chi-square Test that All Eigenvalues are Equal

N : 49.0000
Chi-square : 1542.2903
df : 78.0000
p-value : 0.0000

Chi-square Test that the Last 10 Eigenvalues are Equal

Chi-square : 636.4350
df : 59.8885
p-value : 0.0000

Component Loadings

	1	2	3
GDP_CAP	0.9769	-0.0366	-0.0606
GNP_86	0.9703	-0.0846	0.0040
BIRTH_RT	-0.9512	0.0136	-0.0774
LITERACY	0.8972	-0.1008	0.3004
EDUC	0.8927	-0.0857	-0.2296
MIL	0.8770	0.1501	-0.2909
URBAN	0.8393	0.1425	0.2300
B_TO_D	-0.5166	-0.1225	0.7762
POP_1990	0.0382	0.9972	0.0394
POP_1986	0.0636	0.9966	0.0253
POP_1983	0.0945	0.9940	0.0248

```
POP_2020  | -0.1796    0.9748    0.1002
DEATH_RT  | -0.4533    0.0820   -0.8662
```

Variance Explained by Components

```
1         2         3
-------------------------
6.3950    4.0165    1.6557
```

Percent of Total Variance Explained

```
1          2          3
----------------------------
49.1924    30.8964    12.7361
```

Rotated Pattern Matrix (OBLIMIN, Gamma = 0.0000)

```
           |       1          2          3
-----------+----------------------------
GDP_CAP    |  0.9779   -0.0399    0.0523
GNP_86     |  0.9714   -0.0816   -0.0146
BIRTH_RT   | -0.9506    0.0040    0.0843
EDUC       |  0.8961   -0.1049    0.2194
LITERACY   |  0.8956   -0.0700   -0.3112
MIL        |  0.8777    0.1242    0.2924
URBAN      |  0.8349    0.1658   -0.2285
B_TO_D     | -0.5224   -0.0501   -0.7787
POP_1990   |  0.0236    0.9977    0.0095
POP_1986   |  0.0491    0.9958    0.0234
POP_1983   |  0.0801    0.9932    0.0235
POP_2020   | -0.1945    0.9805   -0.0510
DEATH_RT   | -0.4459   -0.0011    0.8730
```

"Variance" Explained by Rotated Components

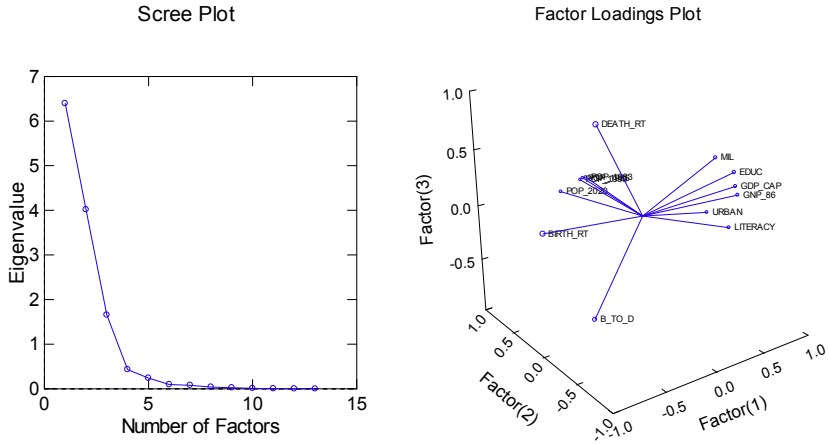
```
1         2         3
-------------------------
6.3946    4.0057    1.6669
```

Percent of Total Variance Explained

```
1          2          3
----------------------------
49.1895    30.8129    12.8225
```

Correlations Among Oblique Factors or Components

```
     1         2         3
------------------------------
1    1.0000
2    0.0127    1.0000
3   -0.0020    0.0452    1.0000
```



By default, SYSTAT extracts three factors because three eigenvalues are greater than 1.0. On factor 1, seven or eight variables have high loadings. The eighth, *B_TO_D* (a ratio of birth-to-death rate) has a higher loading on factor 3. With the exception of *BIRTH_RT*, the other variables are economic measures, so let us identify this as the "economic" factor. Clearly, the second factor can be named "population," and the third, less clearly, "death rates."

The economic and population factors account for 80% (49.19 + 30.81) of the total variance, so a plot of the scores for these factors should be useful for characterizing differences among the countries. The third factor accounts for 13% of the total variance, a much smaller amount than the other two factors. Notice, too, that only 7% of the total variance is not accounted for by these three factors.

Revisiting the Correlation Matrix

Let us examine the correlation matrix for these variables. In an effort to group the variables contributing to each factor, we order the variables according to their factor loadings for the factor on which they load the highest.

The input is:

```
CORR
   USE OURWORLD
   LET (GDP_CAP, GNP_86, POP_1983, POP_1986, POP_1990,
        POP_2020) = L10(@)
   LET (MIL,EDUC) = SQR(@)
   PEARSON GDP_CAP GNP_86 BIRTH_RT EDUC LITERACY MIL URBAN ,
           POP_1990 POP_1986 POP_1983 POP_2020 B_TO_D DEATH_RT
```

The output is:

Pearson Correlation Matrix

	GDP_CAP	GNP_86	BIRTH_RT	EDUC	LITERACY	MIL
GDP_CAP	1.0000					
GNP_86	0.9736	1.0000				
BIRTH_RT	-0.9189	-0.8786	1.0000			
EDUC	0.8996	0.9207	-0.7528	1.0000		
LITERACY	0.8337	0.8404	-0.9302	0.6869	1.0000	
MIL	0.8657	0.8514	-0.7547	0.8869	0.6421	1.0000
URBAN	0.7636	0.7747	-0.8002	0.6238	0.7997	0.6453
POP_1990	-0.0015	-0.0447	-0.0252	-0.0513	-0.0534	0.1727
POP_1986	0.0248	-0.0215	-0.0523	-0.0306	-0.0327	0.1942
POP_1983	0.0583	0.0090	-0.0836	-0.0062	-0.0050	0.2206
POP_2020	-0.2116	-0.2484	0.1880	-0.2555	-0.2360	-0.0339
B_TO_D	-0.5293	-0.4411	0.5106	-0.5252	-0.2737	-0.6184
DEATH_RT	-0.4012	-0.4518	0.5110	-0.2151	-0.6601	-0.1482

	URBAN	POP_1990	POP_1986	POP_1983	POP_2020	B_TO_D
GDP_CAP						
GNP_86						
BIRTH_RT						
EDUC						
LITERACY						
MIL						
URBAN	1.0000					
POP_1990	0.1700	1.0000				
POP_1986	0.1898	0.9992	1.0000			
POP_1983	0.2133	0.9966	0.9984	1.0000		
POP_2020	0.0054	0.9673	0.9605	0.9531	1.0000	
B_TO_D	-0.3074	-0.1070	-0.1358	-0.1526	0.0617	1.0000
DEATH_RT	-0.5126	0.0284	0.0291	0.0152	0.0743	-0.4340

Pearson Correlation Matrix (contd...)

	DEATH_RT
GDP_CAP	
GNP_86	
BIRTH_RT	
EDUC	
LITERACY	
MIL	
URBAN	
POP_1990	
POP_1986	
POP_1983	
POP_2020	
B_TO_D	
DEATH_RT	1.0000

The top triangle of the matrix shows the correlations of the variables within the "economic" factor. *BIRTH_RT* has strong negative correlations with the other variables. Correlations of the population variables with the economic variables are displayed in the four rows below this top portion, and correlations of the death rates variables with the economic variables are in the next two rows. Correlations within the population factor are displayed in the top triangle of the bottom panel. The correlation between the variables in factor 3 (*B_TO_D* and *DEATH_RT*) is –0.434 and is smaller than any of the other within-factor correlations.

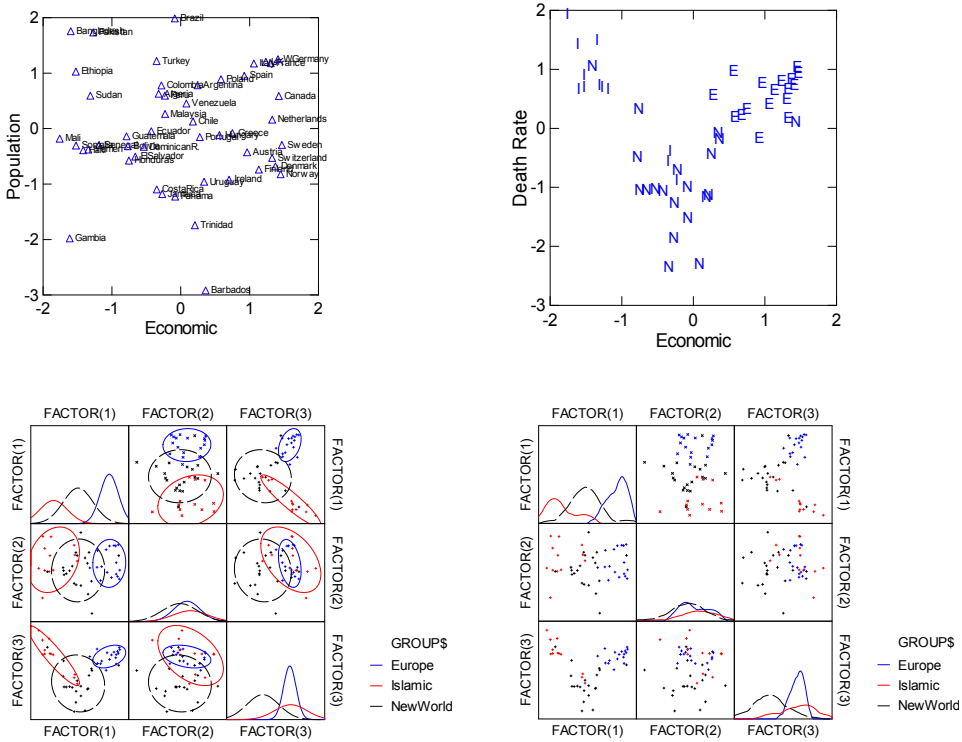
Factor Scores

Look at the scores just stored in *PCASCORE*. First, merge the name of each country and the grouping variable *GROUP$* with the scores. The values of *GROUP$* identify each country as Europe, Islamic, or New World. Next, plot factor 2 against factor 1 (labeling points with country names) and factor 3 against factor 1 (labeling points with the first letter of their group membership). Finally, use SPLOMs to display the scores, adding 75% confidence ellipses for each subgroup in the plots and normal curves for the univariate distributions. Repeat the latter using kernel density estimators.

The input is:

```
MERGE PCASCORE.SYD(FACTOR(1) FACTOR(2) FACTOR(3)),
OURWORLD.SYD(GROUP$ COUNTRY$)
PLOT FACTOR(2)*FACTOR(1) / XLABEL='Economic' ,
                           YLABEL='Population' SYMBOL=4,2,3,
                          SIZE= 1.250 LABEL=COUNTRY$ CSIZE=1.250
PLOT FACTOR(3)*FACTOR(1) / XLABEL='Economic' ,
                           YLABEL='Death Rate' COLOR=2,1,10,
                         SYMBOL=GROUP$ SIZE= 1.250 ,1.250 ,1.250
SPLOM FACTOR(1) FACTOR(2) FACTOR(3)/ GROUP=GROUP$ OVERLAY,
                                     DENSITY=NORMAL  ELL =0.750,
                                     COLOR=2,1,10  SYMBOL=4,2,3,
                                     DASH=1,1,4
SPLOM FACTOR(1) FACTOR(2) FACTOR(3)/ GROUP=GROUP$ OVERLAY,
                                    DENSITY=KERNEL COLOR=2,1,10,
                                     SYMBOL=4,2,3 DASH=1,1,4
```

The output is:



High loadings on the "economic" factor show countries that are strong economically (Germany, Canada, Netherlands, Sweden, Switzerland, Denmark, and Norway) relative to those with low loadings (Bangladesh, Ethiopia, Mali, and Gambia). Not surprisingly, the population factor identifies Barbados as the smallest and Bangladesh, Pakistan, and Brazil as largest. The questionable third factor (death rate) does help to separate the New World countries from the others.

In each SPLOM, the dashed lines marking curves, ellipses, and kernel contours identify New World countries. The kernel contours in the plot of factor 3 against factor 1 identify a pocket of Islamic countries within the New World group.

Computation

Algorithms

Provisional methods are used for computing covariance or correlation matrices (see Correlations for references). Components are computed by using a Householder tridiagonalization and implicit QL iterations. Rotations are computed with a variant of Kaiser's iterative algorithm, described in Mulaik (1972).

Missing Data

Ordinarily, Factor Analysis and other multivariate procedures delete all cases having missing values on any variable selected for analysis. This is listwise deletion. For data with many missing values, you may end up with too few complete cases for analysis. Select Pairwise deletion if you want covariances or correlations computed separately for each pair of variables selected for analysis. Pairwise deletion takes more time than the standard listwise deletion because all possible pairs of variances and covariances are computed. The same option is offered for Correlations, should you decide to create a symmetric matrix for use in factor analysis that way. Also notice that Correlation provides an EM algorithm for estimating correlation or covariance matrices when data are missing.

Be careful. When you use pairwise deletion, you can end up with negative eigenvalues for principal components or be unable to compute common factors at all. With either method, it is desirable that the pattern of missing data be random. Otherwise, the factor structure you compute will be influenced systematically by the pattern of how values are missing.

References

Afifi, A. A., May, S., and Clark, V. (2004). *Computer-aided multivariate analysis,* 4th ed. New York: Chapman & Hall.

Clarkson, D. B. and Jennrich, R. I. (1988). Quartic rotation criteria and algorithms. *Psychometrika*, 53, 251–259.

Dixon, W. J. (1992). *BMDP statistical software manual*. Berkeley: University of California Press.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. 2nd ed. New York: John Wiley & Sons.

Harman, H. H. (1976). *Modern factor analysis*, 3rd ed. Chicago: University of Chicago Press.

Jackson, J. E. (2003). *A user's guide to principal components*. New ed. New York: Wiley Interscience.

* Jennrich, R.I. and Robinson, S.M. (1969). A Newton-Raphson algorithm for maximum likelihood factor analysis. *Psychometrika, 34, 111-123.*

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. London: Academic Press.

Morrison, D. F. (2004). *Multivariate statistical methods*, 5th ed. CA: Duxbury Press.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Rozeboom, W. W. (1982). The determinacy of common factors in large item domains. *Psychometrika*, 47, 281–295.

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and 1970's: some interesting parallels. *Psychometrika*, 44, 157–167.

(* indicates additional references)