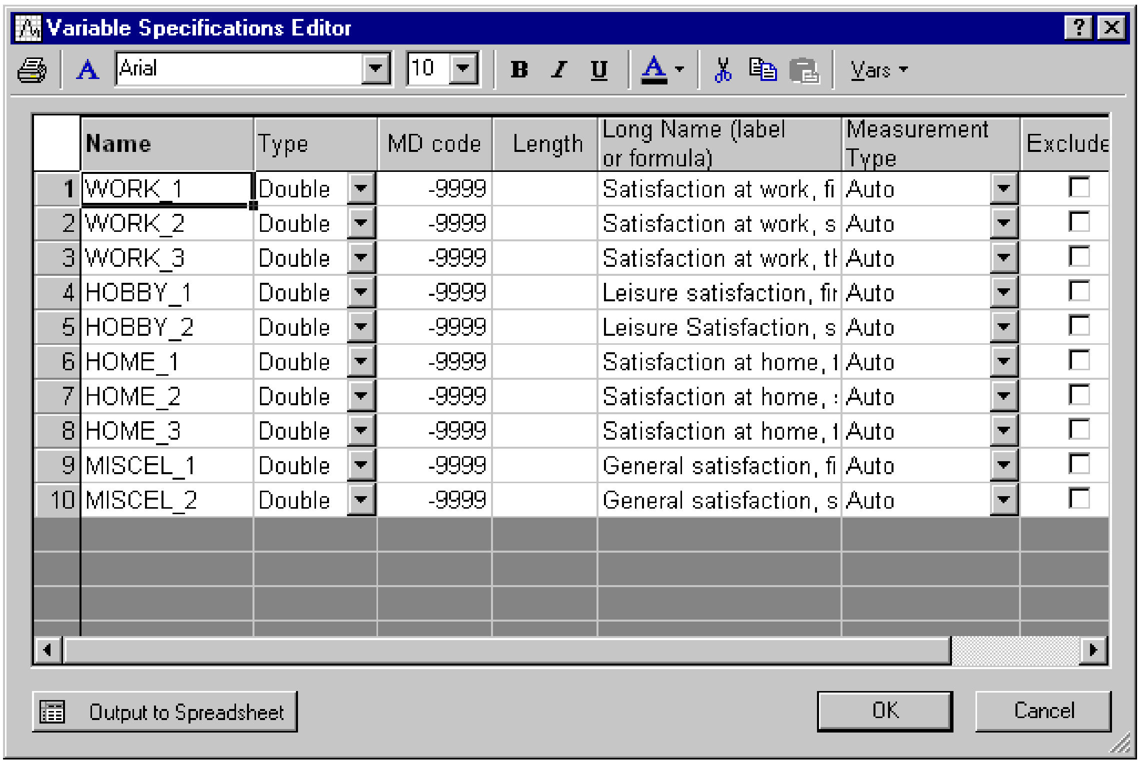
Example 1: Factor Analysis

Overview. This example is based on a fictitious data set describing a study of life satisfaction. Suppose that a questionnaire is administered to a random sample of 100 adults. The questionnaire contains 10 items that are designed to measure satisfaction at work, satisfaction with hobbies, satisfaction at home, and general satisfaction in other areas of life. Responses to all questions are recorded via computer and scaled so that the mean for all items is approximately 100.

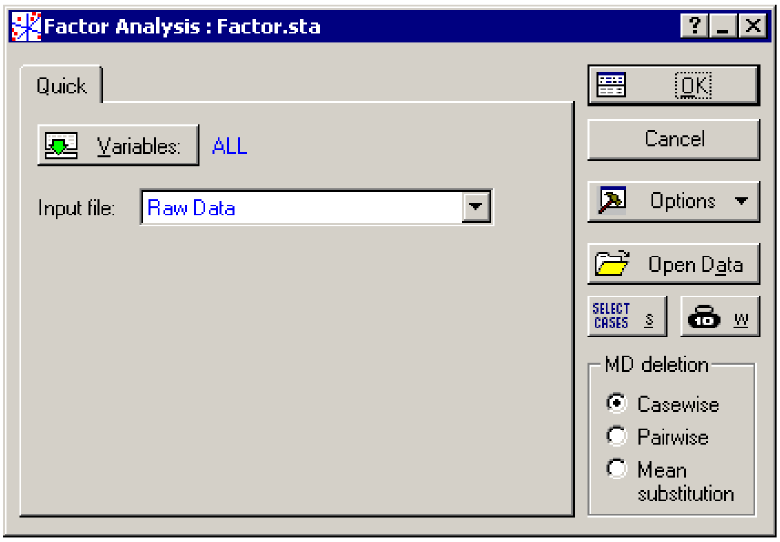
The results for all respondents are entered into the *Factor.sta* data file. Open this data file via the *File* - *Open Examples* menu; it is in the *Datasets* folder. Below is a listing of the variables in that file (to obtain this listing, select *All Variables Specs* from the *Data* menu).



	Name	Type	MD code	Length	Long Name (label or formula)	Measurement Type	Exclude
1	WORK_1	Double	-9999		Satisfaction at work, fi	Auto	
2	WORK_2	Double	-9999		Satisfaction at work, s	Auto	
3	WORK_3	Double	-9999		Satisfaction at work, tl	Auto	
4	HOBBY_1	Double	-9999		Leisure satisfaction, fir	Auto	
5	HOBBY_2	Double	-9999		Leisure Satisfaction, s	Auto	
6	HOME_1	Double	-9999		Satisfaction at home, 1	Auto	
7	HOME_2	Double	-9999		Satisfaction at home, :	Auto	
8	HOME_3	Double	-9999		Satisfaction at home, 1	Auto	
9	MISCEL_1	Double	-9999		General satisfaction, fi	Auto	
10	MISCEL_2	Double	-9999		General satisfaction, s	Auto	

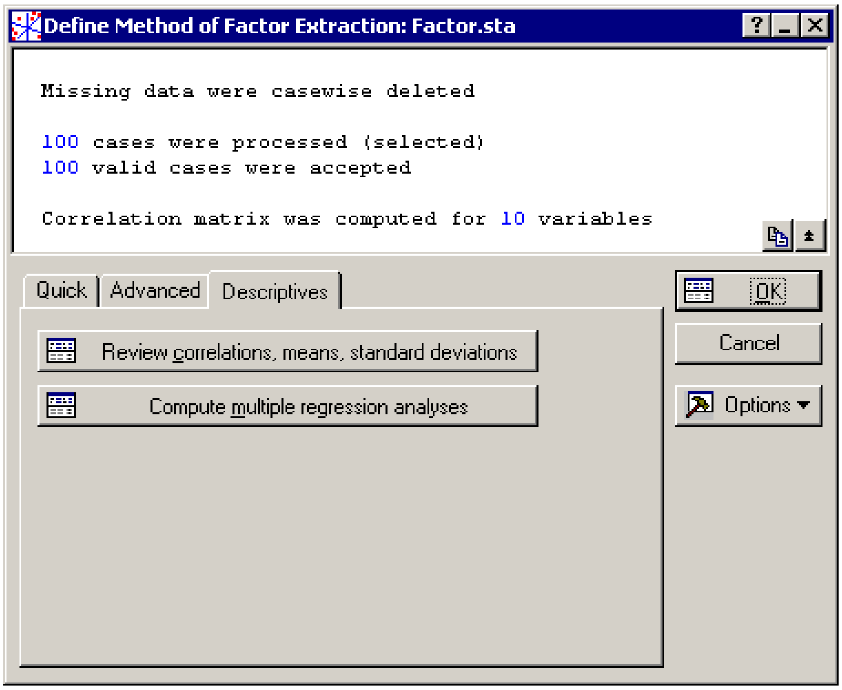
Purpose of analysis. The goal is to learn more about the relationships between satisfaction in the different domains. Specifically, it was desired to learn about the number of factors "behind" these different domains of satisfaction, and their meaning.

Specifying the Analysis. Select *Factor Analysis* from the *Statistics* - *Multivariate Exploratory Analysis* menu to display the *Factor Analysis Startup Panel*. Click the *Variables* button, select all 10 variables, and click the *OK* button. The Startup Panel will look as shown below:

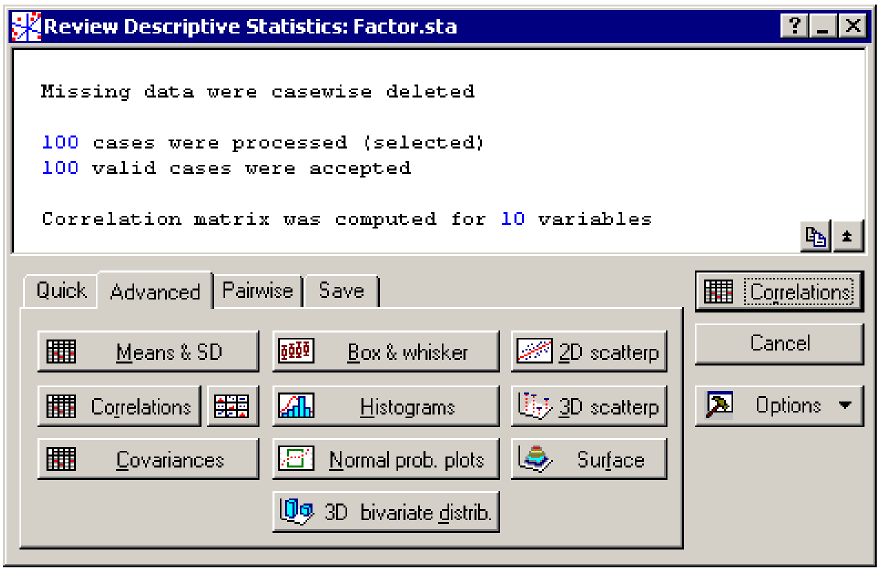


Other options. In order to perform a standard factor analysis, this is all that you need to specify on this dialog. Note that you could also choose either *Casewise* or *Pairwise deletion*, or *Mean substitution* of missing data (via the *MD deletion* group) or a *Correlation Matrix* data file (via the *Input file* box).

Define Method of Factor Extraction. Now, click the *OK* button to display the *Define Method of Factor Extraction* dialog. In this dialog, you can review descriptive statistics, perform a multiple regression analysis, select the extraction method for the factor analysis, select the maximum number of factors and the minimum eigenvalue, and select other options related to specific extraction methods. For now, click on the *Descriptives* tab.



Review Descriptive Statistics. Now, click the *Review correlations, means, standard deviations* button to display the *Review Descriptive Statistics* dialog.



Here, you can review the descriptive statistics graphically or through spreadsheets.

Computing correlation matrix. Click the *Correlations* button on the *Advanced tab* to display the *Correlations* spreadsheet.

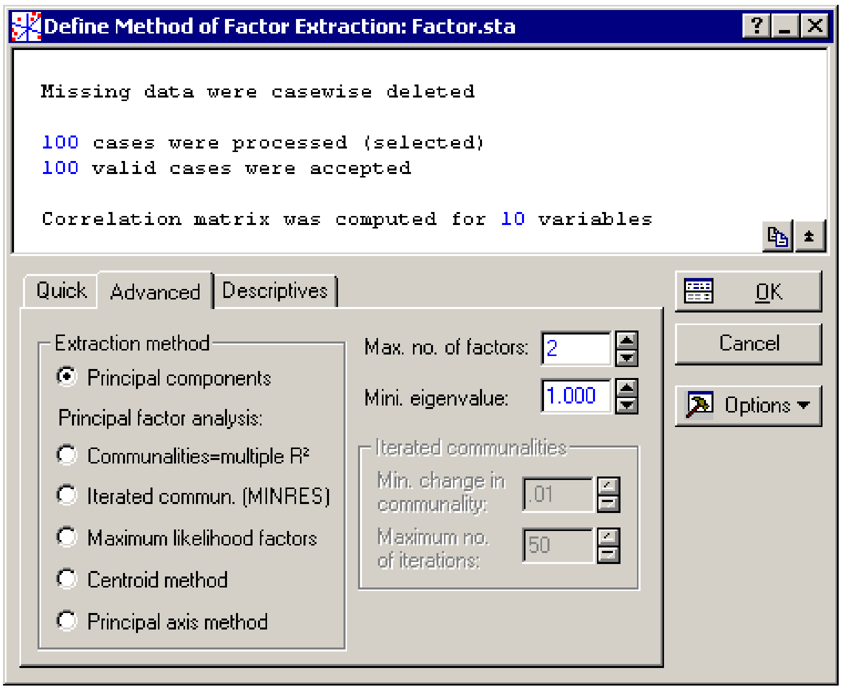
Data: Correlations (Factor.sta)

Correlations (Factor.sta)
Casewise deletion of MD
N=100

Variable	WORK_1	WORK_2	WORK_3	HOBBY_1	HOBBY_2	HOME_1	HOME_2	HOME_3	MISCEL_1	MISCEL_2
WORK_1	1.00	0.65	0.65	0.60	0.52	0.14	0.15	0.14	0.61	0.55
WORK_2	0.65	1.00	0.73	0.69	0.70	0.14	0.18	0.24	0.71	0.68
WORK_3	0.65	0.73	1.00	0.64	0.63	0.16	0.24	0.25	0.70	0.67
HOBBY_1	0.60	0.69	0.64	1.00	0.80	0.54	0.63	0.58	0.90	0.84
HOBBY_2	0.52	0.70	0.63	0.80	1.00	0.51	0.50	0.48	0.81	0.76
HOME_1	0.14	0.14	0.16	0.54	0.51	1.00	0.66	0.59	0.50	0.42
HOME_2	0.15	0.18	0.24	0.63	0.50	0.66	1.00	0.73	0.64	0.59
HOME_3	0.14	0.24	0.25	0.58	0.48	0.59	0.73	1.00	0.59	0.52
MISCEL_1	0.61	0.71	0.70	0.90	0.81	0.50	0.64	0.59	1.00	0.84
MISCEL_2	0.55	0.68	0.67	0.84	0.76	0.42	0.59	0.52	0.84	1.00

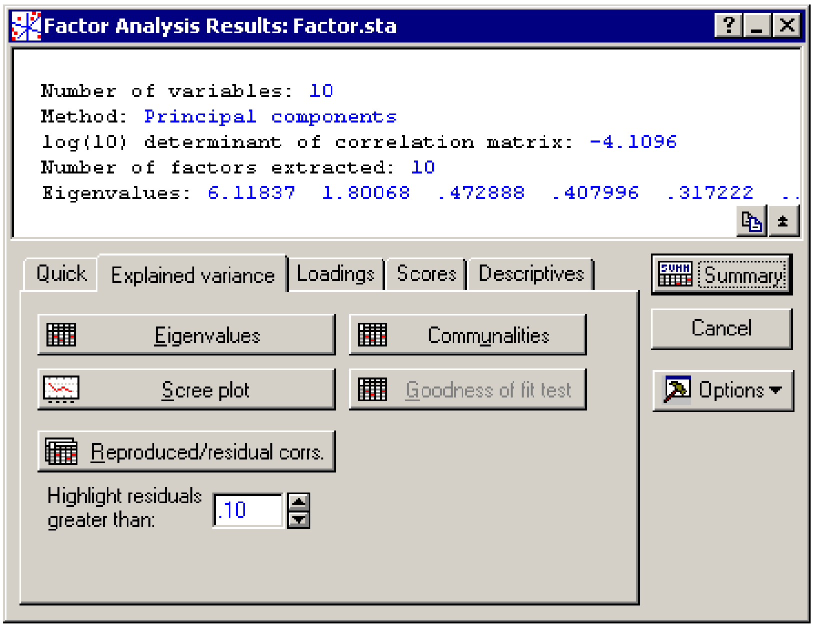
All correlations in the this spreadsheet are positive; some correlations are of substantial magnitude. For example, variables *Hobby_1* and *Miscel_1* are correlated at the level of *.90*. Some correlations (for example the ones between work satisfaction and home satisfaction) seem comparatively small. So, it looks like there is some clear structure in this matrix.

Extraction Method. Now, click the *Cancel* button to return to the *Define Method of Factor Extraction* dialog. You can choose from several extraction methods on the *Advanced* tab (see the *Define Method of Factor Extraction - Advanced* tab topic for a description of each method, and the Introductory Overviews for a description of *Principal Components* and *Principal Factors*). For this example, accept the default extraction method of *Principal Components* and change the *Max. no. of factors* to *10* (the maximum number of factors in this example) and the *Mini. eigenvalue* to *0* (the minimum value for this option).



Click the *OK* button on this dialog to continue the analysis.

Reviewing Results. You can interactively review the results of the factor analysis in the *Factor Analysis Results* dialog. First, click on the *Explained Variance* tab.



Reviewing the eigenvalues. The meaning of eigenvalues and how they help you decide how many factors to retain (interpret) is explained in the Introductory Overviews. Click the *Eigenvalues* button to display the spreadsheet of eigenvalues, percent of total variance, cumulative eigenvalues, and cumulative percent.

Data: Eigenvalues (Factor.sta)

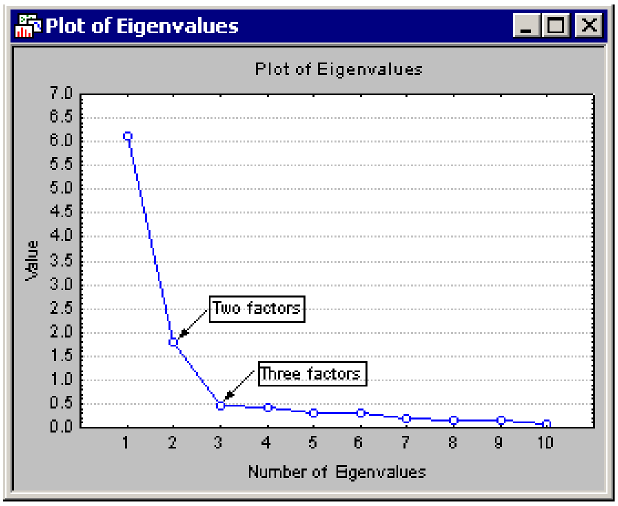
Eigenvalues (Factor.sta)
Extraction: Principal components

Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	6.118369	61.18369	6.11837	61.1837
2	1.800682	18.00682	7.91905	79.1905
3	0.472888	4.72888	8.39194	83.9194
4	0.407996	4.07996	8.79993	87.9993
5	0.317222	3.17222	9.11716	91.1716
6	0.293300	2.93300	9.41046	94.1046
7	0.195808	1.95808	9.60626	96.0626
8	0.170431	1.70431	9.77670	97.7670
9	0.137970	1.37970	9.91467	99.1467
10	0.085334	0.85334	10.00000	100.0000

As you can see, the eigenvalue for the first factor is equal to *6.118369*; the proportion of variance accounted for by the first factor is approximately *61.2%*. Note that these values happen to be easily comparable here because there are 10 variables in the analysis, and thus the sum of all eigenvalues is equal to 10. The second factor accounts for about *18%* of the variance. The remaining eigenvalues each account for less than *5%* of the total variance.

Deciding on the Number of Factors. The Introductory Overviews briefly describe how these eigenvalues can be used to decide how many factors to retain, that is, to interpret. According to the Kaiser criterion (Kaiser, 1960), you would retain factors with an eigenvalue greater than 1. Based on the eigenvalues in the *Eigenvalues* spreadsheet shown above, that criterion would suggest you choose 2 factors.

Scree test. Now, to produce a line graph of the eigenvalues in order to perform Cattell's scree test (Cattell, 1966), click the *Scree plot* button. The *Plot of Eigevalues* graph shown below has been "enhanced" to clarify the test. Based on Monte Carlo studies, Cattell suggests that the point where the continuous drop in eigenvalues levels off suggests the cutoff, where only random "noise" is being extracted by additional factors. In our example, that point could be at factor 2 or factor 3 (as indicated by the arrows). Therefore, you should try both solutions and see which one will yield the most interpretable factor pattern.



Now, examine the factor loadings.

Factor Loadings. As described in the Introductory Overviews, factor loadings can be interpreted as the correlations between the factors and the variables. Thus, they represent the most important information on which the interpretation of factors is based. First look at the (unrotated) factor loadings for all 10 factors. Click on the *Factor Analysis Results dialog - Loadings tab* and select *Unrotated* from the *Factor rotation* box. Then click the *Summary: Factor Loadings* button to display the *Factor Loadings* spreadsheet of loadings.

Data: Factor Loadings (Unrotated) (Factor.sta)

Factor Loadings (Unrotated) (Factor.sta)
Extraction: Principal components
(Marked loadings are > .700000)

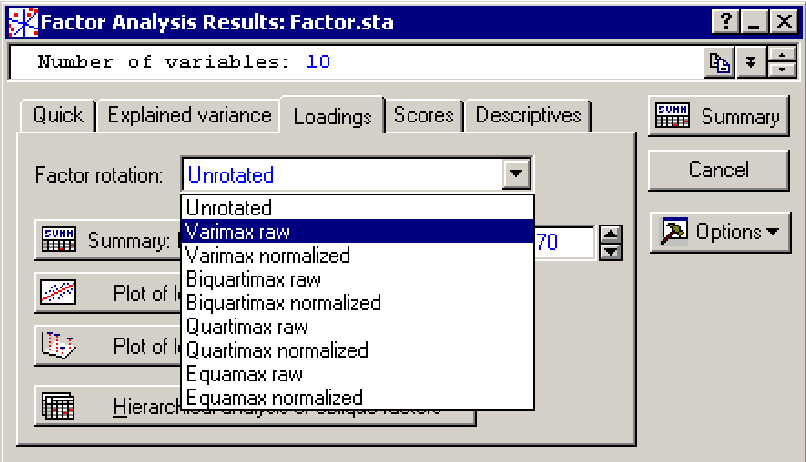
Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 9	Factor 10
WORK_1	-0.652601	0.514217	0.301687	0.439108	0.080008	0.003894
WORK_2	-0.756976	0.494770	-0.078826	-0.211795	0.103633	0.012210
WORK_3	-0.745706	0.456680	-0.104749	0.030826	-0.017932	0.038980
HOBBY_1	-0.941630	-0.021835	0.012653	0.001861	-0.243305	0.171990
HOBBY_2	-0.875615	0.051643	0.099675	-0.324541	0.088684	0.017996
HOME_1	-0.576062	-0.604977	0.490999	-0.114927	0.004027	-0.019576
HOME_2	-0.671289	-0.617962	-0.125776	0.159963	0.145372	0.048318
HOME_3	-0.641532	-0.573925	-0.268572	0.152709	0.006890	0.000902
MISCEL_1	-0.951516	0.013513	-0.050164	0.026706	-0.156713	-0.223847
MISCEL_2	-0.900333	0.048154	-0.151805	-0.034832	0.087690	-0.030324
Expl.Var	6.118369	1.800682	0.472888	0.407996	0.137970	0.085334
Prp.Totl	0.611837	0.180068	0.047289	0.040800	0.013797	0.008533

Remember that factors are extracted so that successive factors account for less and less variance (see the Introductory Overviews). Therefore, it is not surprising to see that the first factor shows most of the highest loadings. Also note that the sign of the factor loadings only counts insofar as variables with opposite loadings on the same factor relate to that factor in opposite ways. However, you could multiply all loadings in a column by -1 (i.e., reverse all signs), and the results would not be affected in any way.

Rotating the Factor Solution. As described in the Introductory Overviews, the actual orientation of the factors in the factorial space is arbitrary, and all rotations of factors will reproduce the correlations equally well. This being the case, it seems natural to rotate the factor solution to yield a factor structure that is simplest to interpret; in fact, the formal term *simple structure* was coined and defined by Thurstone (1947) to basically describe the condition when factors are marked by high loadings for some variables, low loadings for others, and when there are few high cross-loadings, that is, few variables with substantial loadings on more than one factor. The most standard computational method of rotation to bring about simple structure is the *varimax* rotation (Kaiser, 1958); others that have been proposed are *quartimax*, *biquartimax*, and *equamax* (see Harman, 1967) and are implemented in *STATISTICA*.

Specifying a rotation. First, consider the number of factors that you want to rotate, that is, retain and interpret. It was previously decided that two is most likely the appropriate number of factors; however, based on the results of the Scree plot, it was also decided to look at the three factor solution. We will start with three factors. Click the *Cancel* button to return to the *Define Method of Factor Extraction* dialog and change the *Maximum no. of factors* on the *Quick tab* from *10* to *3*. Then click the *OK* button to continue with the analysis.

On the *Factor Analysis Results dialog - Loadings tab*, select *Varimax Raw* from the *Factor rotation* box to perform a varimax rotation.



Click the *Summary: Factor loadings* button to display the *Factor loadings* spreadsheet.

Data: Factor Loadings (Varimax raw) (Factor.sta)*

Factor Loadings (Varimax raw) (Factor.sta)
Extraction: Principal components
(Marked loadings are > .700000)

Variable	Factor 1	Factor 2	Factor 3
WORK_1	0.839579	-0.157384	0.227287
WORK_2	0.898615	0.118837	-0.048899
WORK_3	0.865608	0.151923	-0.057003
HOBBY_1	0.731037	0.501711	0.318078
HOBBY_2	0.726495	0.371499	0.336895
HOME_1	0.099696	0.426704	0.864238
HOME_2	0.148303	0.823540	0.384856
HOME_3	0.147422	0.857607	0.236350
MISCEL_1	0.758585	0.518368	0.252834
MISCEL_2	0.736270	0.524719	0.136161
Expl.Var	4.495100	2.591518	1.305320
Prp.Totl	0.449510	0.259152	0.130532

Reviewing the three-factor rotated solution. In the *Factor loadings* spreadsheet above, substantial loadings on the first factor appear for all but the home-related items. *Factor 2* shows fairly substantial factor loadings for all but the work-related satisfaction items. *Factor 3* only has one substantial loading for variable *Home_1*. The fact that only one variable shows a high loading on the third factor makes one wonder whether one cannot do just as well without it (the third factor).

Reviewing the two-factor rotated solution. Once again, click the *Cancel* button on the *Factor Analysis Results* dialog to return to the *Define Method of Factor Extraction* dialog. Change the *Maximum no. of factors* on the *Quick tab* from *3* to *2* and click the *OK* button to continue to the *Factor Analysis Results* dialog. Again select *Varimax raw* from the *Factor rotation* box on the *Loadings tab* and click the *Summary: Factor loadings* button.

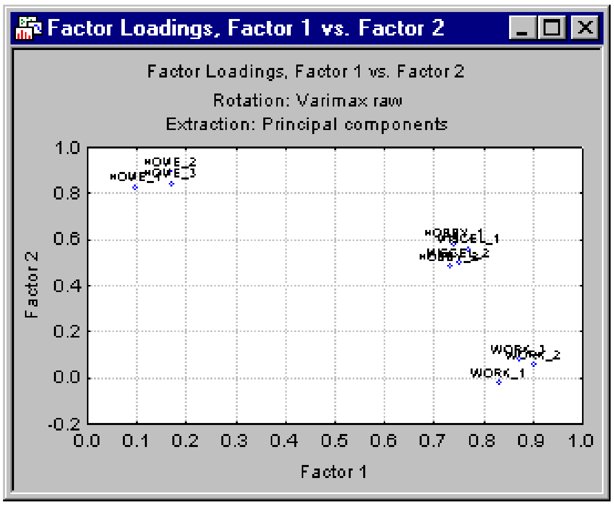
Data: Factor Loadings (Varimax raw) (Factor.st...

Variable	Factor Loadings (Varimax raw) (Factor.sta) Extraction: Principal components (Marked loadings are > .700000) Factor 1	Factor 2
WORK_1	0.830623	-0.019320
WORK_2	0.902408	0.058905
WORK_3	0.870524	0.082595
HOBBY_1	0.739857	0.582885
HOBBY_2	0.731191	0.484489
HOME_1	0.097371	0.829676
HOME_2	0.165722	0.897242
HOME_3	0.168370	0.844159
MISCEL_1	0.768988	0.560555
MISCEL_2	0.748861	0.502121
Expl.Var	4.561544	3.357507
Prp.Totl	0.456154	0.335751

Factor 1 shows the highest loadings for the items pertaining to work-related satisfaction. The smallest loadings on that factor are for home-related satisfaction items. The other loadings fall in-between. *Factor 2* shows the highest loadings for the home-related satisfaction items, lowest loadings for work-related satisfaction items, and loadings in-between for the other items.

Interpreting the two-factor rotated solution. Does this pattern lend itself to an easy interpretation? It looks like the two factors are best identified as the work satisfaction factor (*Factor 1*) and the home satisfaction factor (*Factor 2*). Satisfaction with hobbies and miscellaneous other aspects of life seem to be related to both factors. This pattern makes some sense in that satisfaction at work and at home may be independent from each other in this sample, but both contribute to leisure time (hobby) satisfaction and satisfaction with other aspects of life.

Plot of the two-factor rotated solution. Click the *Plot of loadings, 2D* button on the *Factor Analysis Results* dialog - *Loadings* tab to produce a scatterplot of the two factors. The graph below simply shows the two loadings for each variable. Note that this scatterplot nicely illustrates the two independent factors and the 4 variables (*Hobby_1*, *Hobby_2*, *Miscel_1*, *Miscel_2*) with the cross-loadings.



Now, see how well you can reproduce the observed correlation matrix from the two-factor solution.

Reproduced and Residual Correlation Matrix. Click the *Reproduced/residual corrs.* button on the *Explained variance* tab to display two spreadsheets with the reproduced correlation matrix and the residual correlations (observed minus reproduced correlations).

Data: Residual Correlations (Factor)*

Residual Correlations (Factor.sta)
Extraction: Principal components
(Marked residuals are > .100000)

Variable	WORK_1	WORK_2	WORK_3	HOBBY_1	HOBBY_2	HOME_1	HOME_2	HOME_3	MISCEL_1	MISCEL_2
WORK_1	0.31	-0.10	-0.07	-0.01	-0.08	0.08	0.02	0.01	-0.02	-0.06
WORK_2	-0.10	0.18	-0.06	-0.01	0.01	0.01	-0.02	0.03	-0.02	-0.02
WORK_3	-0.07	-0.06	0.24	-0.06	-0.05	0.01	0.02	0.04	-0.02	-0.02
HOBBY_1	-0.01	-0.01	-0.06	0.11	-0.02	-0.02	-0.01	-0.03	0.01	-0.00
HOBBY_2	-0.08	0.01	-0.05	-0.02	0.23	0.03	-0.06	-0.05	-0.02	-0.04
HOME_1	0.08	0.01	0.01	-0.02	0.03	0.30	-0.10	-0.13	-0.04	-0.06
HOME_2	0.02	-0.02	0.02	-0.01	-0.06	-0.10	0.17	-0.05	0.01	0.02
HOME_3	0.01	0.03	0.04	-0.03	-0.05	-0.13	-0.05	0.26	-0.02	-0.03
MISCEL_1	-0.02	-0.02	-0.02	0.01	-0.02	-0.04	0.01	-0.02	0.09	-0.02
MISCEL_2	-0.06	-0.02	-0.02	-0.00	-0.04	-0.06	0.02	-0.03	-0.02	0.19

The entries in the *Residual Correlations* spreadsheet can be interpreted as the "amount" of correlation that cannot be accounted for with the two factor solution. Of course, the diagonal elements in the matrix contain the standard deviation that cannot be accounted for, which is equal to the square root of one minus the respective communalities for two factors (remember that the communality of a variable is the variance that can be explained by the respective number of factors). If you review this matrix carefully you will see that there are virtually no residual correlations left that are greater than *0.1* or less than *-0.1* (actually, a few are about of that magnitude). Added to that is the fact that the first two factors accounted for *79%* of the total variance (see *Cumulative %* eigenvalues displayed in the *Eigenvalues* spreadsheet).

The "Secret" to the Perfect Example. The example you have reviewed does indeed provide a nearly perfect two-factor solution. It accounts for most of the variance, allows for ready interpretation, and reproduces the correlation matrix with only minor disturbances (remaining residual correlations). Of course, nature rarely affords one such simplicity, and, indeed, this fictitious data set was generated via the normal random number generator accessible in the spreadsheet formulas. Specifically, two orthogonal (independent) factors were "planted" into the data, from which the correlations between variables were generated. The factor analysis example retrieved those two factors as intended (i.e., the work satisfaction factor and the home satisfaction factor); thus, had nature planted the two factors, you would have learned something about the underlying or latent structure of nature.

Miscellaneous Other Results. Before concluding this example, brief comments on some other results will be made.

Communalities. To view the communalities for the current solution, that is, current numbers of factors, click the *Communalities* button on the *Factor Analysis Results* dialog - *Explained Variance* tab. Remember that the communality of a variable is the portion that can be reproduced from the respective number of factors; the rotation of the factor space has no bearing on the communalities. Very low communalities for one or two variables (out of many in the analysis) may indicate that those variables are not well accounted for by the respective factor model.

Factor score coefficients. The factor score coefficients can be used to compute factor scores. These coefficients represent the weights that are used when computing factor scores from the variables. The coefficient matrix itself is usually of little interest; however, factor scores are useful if one wants to perform further analyses on the factors. To view these coefficients, click the *Factor score coefficients* button on the *Factor Analysis Results* dialog - *Scores* tab.

Factor scores. Factor scores (values) can be thought of as the actual values for each respondent on the underlying factors that you discovered. Click the *Factor scores* button on the *Factor Analysis Results* dialog - *Scores* tab to compute factor scores. These scores can be saved via the *Save factor scores* button and used later in other data analyses.

Final Comment. Factor analysis is a not a simple procedure. Anyone who is routinely using factor analysis with many (e.g., 50 or more) variables has seen a wide variety of "pathological behaviors" such as negative eigenvalues, uninterpretable solutions, ill-conditioned matrices, and such adverse conditions. If you are interested in using factor analysis in order to detect structure or meaningful factors in large numbers of variables, it is recommended that you carefully study a textbook on the subject (such as Harman, 1968). Also, because many crucial decisions in factor analysis are by nature subjective (number of factors, rotational method, interpreting loadings), be prepared for the fact that experience is required before you feel comfortable making those judgments. The *Factor Analysis* module of *STATISTICA* was specifically designed to make it easy for you to switch interactively between different numbers of factors, rotations, etc., so that different solutions can be tried and compared.