Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Principal component analysis attempts to express a n by m data set with $m > 2$ in new coordinates Y obtained by rotating the original coordinates X so that the overall variance of the observations is contained in decreasing order in the new variables. If this is successful in that most of the variance is contained in the first 2 or 3 of the Y variables, then this allows inferences to be drawn about the data in a sub-space of dimension $< m$.

Example 1

From the main SIMFIT menus choose [Statistics], [Multivariate], then [Principal components] and analyze the default test file provided (`g03aaf.tf1`) which contains the following data

7	4	3
4	1	8
6	3	5
8	6	1
8	5	7
7	2	9
5	3	3
9	5	8
7	4	5
8	2	2

leading to these results.

Variables included: 1 2 3
Transformation: Untransformed
Matrix type: Variance-covariance matrix
Score type: Score variance = eigenvalue
Replicates: Unweighted for replicates

Eigenvalues	Proportion	Cumulative	χ^2	DOF	p
8.274	0.6515	0.6515	8.613	5	0.1255
3.676	0.2895	0.9410	4.118	2	0.1276
0.750	0.0590	1.0000	0.000	0	0.0000

Loadings (by column)

-0.138	0.699	0.702
-0.250	0.661	-0.707
0.958	0.273	-0.084

Scores (by column)

-2.150	-0.173	-0.107
3.800	-2.890	-0.510
0.153	-0.987	-0.269
-4.710	1.300	-0.652
1.290	2.280	-0.449
4.100	0.144	0.803
-1.630	-2.230	-0.803
2.110	3.250	0.168
-0.235	0.373	-0.275
-2.750	-1.070	2.090

The significance of the options used for the analysis and the results listed in this table are now explained.

- **Variables included**
  All three variables were included as this was defined in the trailer section of `g03aaf.tf1`, but the variables to be included can also be adjusted interactively.

- **Transformation**
  The data were used without any transformation.

- **Matrix type**
  If the magnitude of the variables are similar so that the data do not need to be centralized and scaled, then the covariance matrix can be used. Otherwise the correlation matrix should be used.

- **Score type**
  Several options are available, to provide consistency and facilitate comparison with published data.

- **Replicates**
  SimFIT provides the facility to supply a weighting vector to permit data suppression (setting a weight to zero), or to allow for replicates (setting a weight equal to the number of replicates used in the observation).

- **Eigenvalues**
  These are listed in decreasing order, the proportion of variance and cumulative sum of variances captured by each component is listed, and a chi-square test is performed to check the significance of each component. The significance levels are not valid if the correlation matrix is used instead of the covariance matrix. Clearly the first two principal components are sufficient to represent the three original variables.

- **Loading**
  Column j of the loading matrix contains the coefficients required to express y_j as linear function of the variables $x_1, x_2, \ldots, x_m$. The values can be used to indicate the importance of the contribution of the original variables to the rotated variables.

- **Scores**
  Row i of the scores matrix contains the values for row i of the original data expressed in variables $y_1, y_2, \ldots, y_m$. If most of the variance can be explained by the first two or three variables, the values can be used instead of the original observations to visualize grouping and clustering, etc.

Example 2

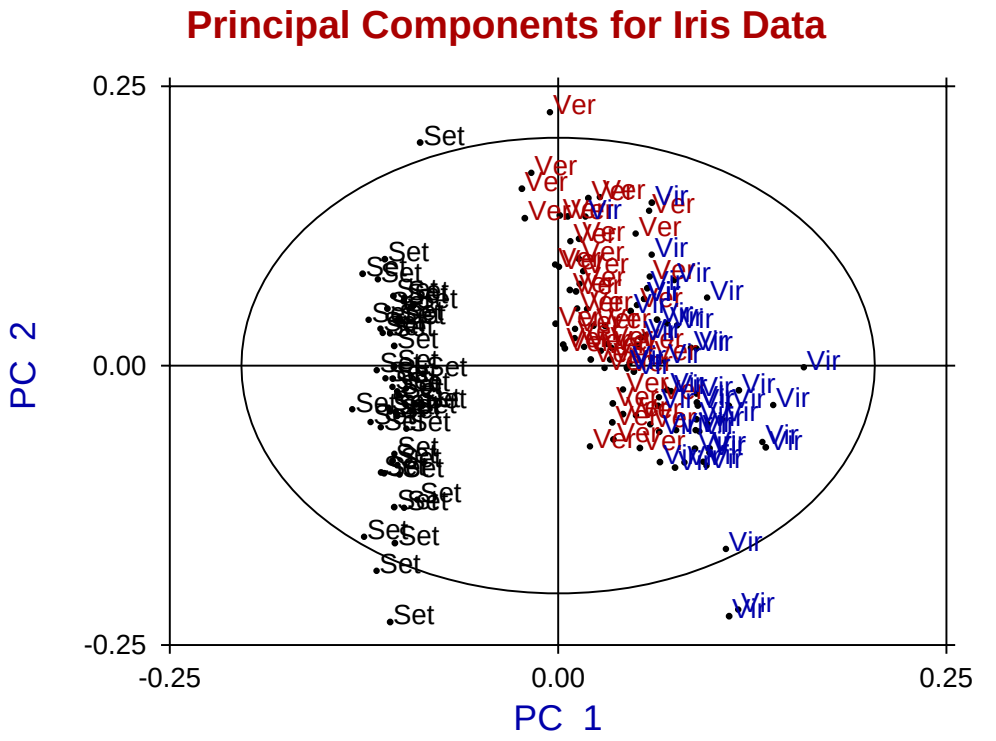
The next example concerns the analysis of the Fisher iris data with 150 cases and 4 variables (Sepal length, Sepal width, Petal length, and Petal width) contained in the test file `iris.tf1`. The figures below show the scores and loadings for these data after analyzing the correlation matrix.

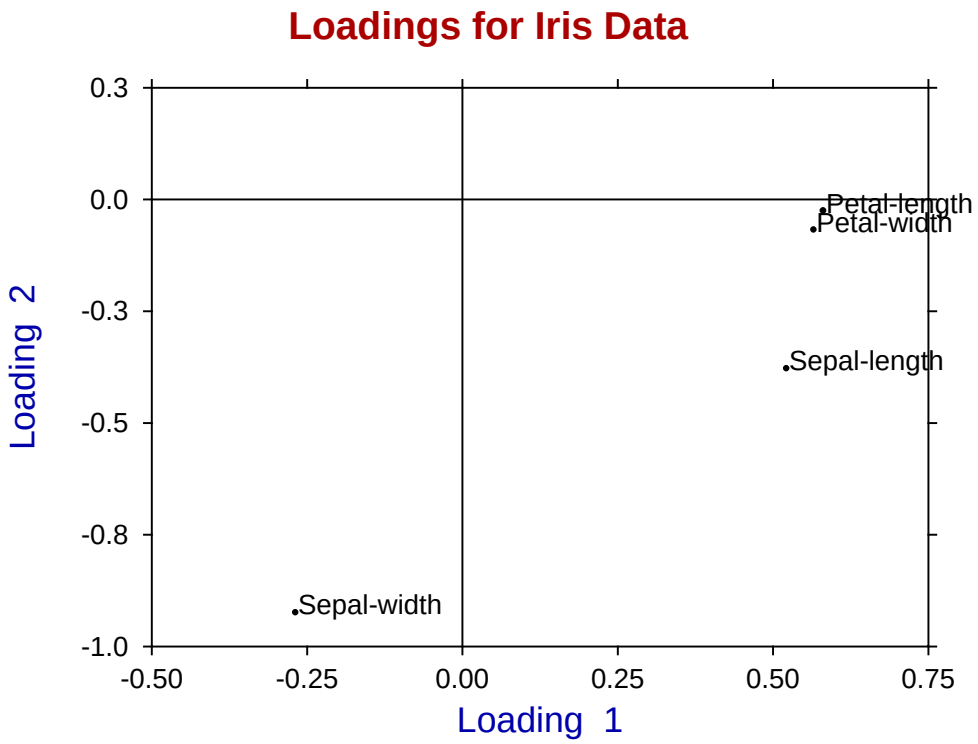
The score plot displays the score components for all samples using the selected principal components, so some may prefer to label the legends as principal components instead of scores, and this plot is used to search for possible groupings among the sample. The components can be labeled using any labels supplied at the end of the data file, but this can cause confusion where, as in the present case, the labels overlap leading to crowding. A method for moving labels to avoid such confusion is provided. However, with such dense labels it is best to just plot the scores using different symbols and colors for the three groups.

The loading plot displays the coefficients that express the selected principal components y_j as linear functions of the original variables $x_1, x_2, \ldots, x_m$, so this plot is used to observe the contributions of the original variables x to the new ones y.

Note that figures also illustrate an application of the SimFIT technique for adding extra data interactively to create the cross-hairs intersecting at $(0, 0)$, and it also shows how labels can be added to identify the variables in a loadings plot. It should be noted that, as the eigenvectors are of indeterminate sign and only the relative

magnitudes of coefficients are important, the scattergrams can be plotted with either the scores calculated from the SVD, or else with the scores multiplied by minus one, which is equivalent to reversing the direction of the corresponding axis in a scores or loadings plot.





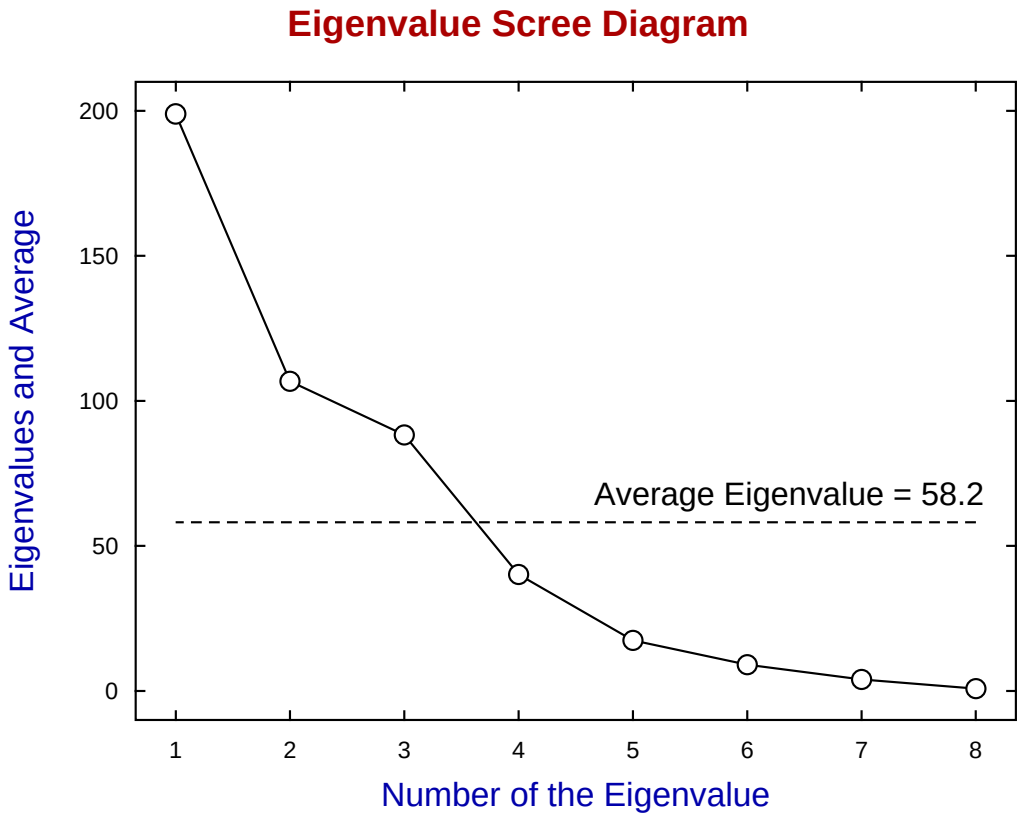
The confidence ellipse plotted on the scores will be explained later.

Example 3

An important topic in principal component analysis is deciding how to choose a sufficient number of principal components to represent the data adequately. As the eigenvalues are proportional to the fractions of variance along the principal component axes, a table of the cumulative proportions is calculated, and some users may find it useful to include sufficient principal components to account for a given amount of the variance, say 70%. Consider these results from the analysis of data with eight variables contained in test file `cluster.tf1` and analyzed using the covariance matrix.

Eigenvalues	Proportion	Cumulative	χ^2	DOF	p
198.9	0.4274	0.4274	61.13	35	0.0041
106.8	0.2294	0.6568	48.09	27	0.0075
88.29	0.1897	0.8465	39.73	20	0.0054
40.18	0.0863	0.9328	25.39	14	0.0309
17.46	0.0375	0.9703	15.04	9	0.0898
9.036	0.0194	0.9898	9.149	5	0.1033
3.966	0.0085	0.9983	4.349	2	0.1137
0.803	0.0017	1.0000	0.000	0	0.0000

The next figure shows how scree plots



can be displayed to illustrate the number of components needed to represent the data adequately. For instance, in this case, it seems that approximately three of the principal components are required. A useful rule of thumb for selecting the minimum number of components is to observe where the scree diagram crosses the average eigenvalue or becomes flattened indicating that all subsequent eigenvalues contribute to a comparable extent. Use of the chi-square statistics for this type of investigation will be described later.

Theory 1: The calculation of principal components

In the principal components analysis of a n by m data matrix, new coordinates y are selected by rotation of the original coordinates x so that the proportion of the variance projected onto the new axes decreases in the order $y_1, y_2, \ldots, y_m$. The hope is that most of the variance can be accounted for by a subset of the data in y coordinates, so reducing the number of dimensions required for data analysis.

It is possible to scale the original data so that the variables are all of comparable dimensions and have similar variances in order to prevent the analysis being dominated by variables with large values. However, basing principal components analysis on the correlation matrix rather than the covariance or sum of squares and cross product matrices is often recommended, as it prevents the analysis being unduly dominated by variables with large values, and is equivalent to centering and scaling the original data. The data format for principal components analysis is exactly the same as for cluster analysis; namely a data matrix with n rows (cases) and m columns (variables).

If the data matrix is X with covariance, correlation or scaled sum of squares and cross products matrix S, then the quadratic form

$$a_1^T S a_1$$

is maximized subject to the normalization $a_1^T a_1 = 1$ to give the first principal component

$$c_1 = \sum_{i=1}^{m} a_{1i} x_i.$$

Similarly, the quadratic form

$$a_2^T S a_2$$

is maximized, subject to the normalization and orthogonality conditions $a_2^T a_2 = 1$ and $a_2^T a_1 = 0$, to give the second principal component

$$c_2 = \sum_{i=1}^{m} a_{2i} x_i$$

and so on. The vectors a_i are the eigenvectors of S with eigenvalues λ_i^2, where the proportion of the variation accounted for by the ith principal component can be estimated as

$$\lambda_i^2 / \sum_{j=1}^{m} \lambda_j^2.$$

Actually SIMFIT uses a singular value decomposition (SVD) of a centered and scaled data matrix, say $X_s = (X - \bar{X})/\sqrt{(n-1)}$ as in

$$X_s = V \Lambda P^T$$

to obtain the diagonal matrix Λ of singular values, the matrix of left singular vectors V as the n by m matrix of scores, and the matrix of right singular vectors P as the m by m matrix of loadings.

Theory 2: Confidence ellipses in scores plots

Note that a 95% confidence Hotelling T^2 ellipse is also plotted, which assumes a multivariate normal distribution for the original data and uses the F distribution.

The confidence ellipse is based on the fact that, if $\bar{y}$ and S are the estimated mean vector and covariance matrix from a sample of size n and, if x is a further independent sample from an assumed $p-$variate normal distribution, then

$$(x - \bar{y})^T S^{-1} (x - \bar{y}) \sim \frac{p(n^2-1)}{n(n-p)} F_{p,n-p},$$

where the significance level for the confidence region can be altered interactively.

Theory 3: The chi-square test for significant components

In cases where the correlation matrix is not used, a chi-square test statistic is also provided along with appropriate probability estimates to make the decision more objective. In this case, if k principal components are selected, the chi-square statistic

$$(n-1-(2m+5)/6)\left\{-\sum_{i=k+1}^{m}\log(\lambda_i^2)+(m-k)\log\left(\sum_{i=k+1}^{m}\lambda_i^2/(m-k)\right)\right\}$$

with $(m-k-1)(m-k+2)/2$ degrees of freedom can be used to test for the equality of the remaining $m-k$ eigenvalues.

If one of these test statistics, say the $k+1$th, is not significant then it is usual to assume k principal components should be retained and the rest regarded as of little importance. So, if it is concluded that the remaining eigenvalues are of comparable importance, then a decision has to be made whether to eliminate all or preserve all. For instance, from the last column of p values referring to the above chi-square test for `g03aaf.tf1`, it might be concluded that a minimum of two components are required to represent this data set adequately. However, for the case of `iris.tf1`, three components would be required.

The common practise of always using two or three components just because these can be visualized is to be deplored.

Theory 4: Calculating scores from loadings

The data used by SimFIT are automatically centered at run time, and sometimes also scaled if requested, so it is not usually necessary to transform the original data for principal component analysis, especially if the correlation matrix method is used. However, in order to calculate scores using the loadings retrospectively the following points should be noted.

1. The original data matrix must be centralized by subtracting column sample means.
2. If the correlation matrix technique was used to calculate the scores, then the data must also be scaled by dividing columns by the column sample standard deviations.
3. If the covariance matrix technique was used no further scaling is required.
4. If the sum of squares and cross-product matrix method was used, then the centralized data must also be multiplied by $\sqrt{n-1}$.
5. The final scaling of the scores will be that used when generating the loadings.
6. The average of a group of k scores is the same as using loadings with the means from the same k values.
7. The scores are unspecified up to multiples of -1.

To illustrate this procedure consider the following steps that are required to calculate the scores for a covariance matrix with scores normalized to have variance equal to the corresponding eigenvalue, using the notation for subroutine `g03aaf` in the NAG library documentation.

- Obtain the data matrix X
- Transform X to obtain the centered matrix Y
- Generate the loading matrix P
- Calculate the scores $V = YP$ as shown next.

$$X = \begin{pmatrix} 7.0 & 4.0 & 3.0 \\ 4.0 & 1.0 & 8.0 \\ 6.0 & 3.0 & 5.0 \\ 8.0 & 6.0 & 1.0 \\ 8.0 & 5.0 & 7.0 \\ 7.0 & 2.0 & 9.0 \\ 5.0 & 3.0 & 3.0 \\ 9.0 & 5.0 & 8.0 \\ 7.0 & 4.0 & 5.0 \\ 8.0 & 2.0 & 2.0 \end{pmatrix}$$

$$Y = \begin{pmatrix} 0.1 & 0.5 & -2.1 \\ -2.9 & -2.5 & 2.9 \\ -0.9 & -0.5 & -0.1 \\ 1.1 & 2.5 & -4.1 \\ 1.1 & 1.5 & 1.9 \\ 0.1 & -1.5 & 3.9 \\ -1.9 & -0.5 & -2.1 \\ 2.1 & 1.5 & 2.9 \\ 0.1 & 0.5 & -0.1 \\ 1.1 & -1.5 & -3.1 \end{pmatrix}$$

$$P = \begin{pmatrix} -0.1376 & 0.6990 & 0.7017 \\ -0.2505 & 0.6609 & -0.7075 \\ 0.9583 & 0.2731 & -0.0842 \end{pmatrix}$$

$$V = YP$$

$$= \begin{pmatrix} 0.1 & 0.5 & -2.1 \\ -2.9 & -2.5 & 2.9 \\ -0.9 & -0.5 & -0.1 \\ 1.1 & 2.5 & -4.1 \\ 1.1 & 1.5 & 1.9 \\ 0.1 & -1.5 & 3.9 \\ -1.9 & -0.5 & -2.1 \\ 2.1 & 1.5 & 2.9 \\ 0.1 & 0.5 & -0.1 \\ 1.1 & -1.5 & -3.1 \end{pmatrix} \begin{pmatrix} -0.1376 & 0.6990 & 0.7017 \\ -0.2505 & 0.6609 & -0.7075 \\ 0.9583 & 0.2731 & -0.0842 \end{pmatrix}$$

$$= \begin{pmatrix} -2.1514 & -0.1731 & -0.1068 \\ 3.8042 & -2.8875 & -0.5104 \\ 0.1532 & -0.9869 & -0.2694 \\ -4.7065 & 1.3015 & -0.6517 \\ 1.2938 & 2.2791 & -0.4492 \\ 4.0993 & 0.1436 & 0.8031 \\ -1.6258 & -2.2321 & -0.8028 \\ 2.1145 & 3.2512 & 0.1684 \\ -0.2348 & 0.3730 & -0.2751 \\ -2.7464 & -1.0689 & 2.0940 \end{pmatrix}$$

Note that, as the sign of eigenvectors is arbitrary and can change with relatively small perturbations of a data set, SimFIT provides the option to reflect plots of loadings and scores in order to retain consistency of spatial distribution for the visual presentations of results.