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

Biplots are widely used to view multivariate data in a space with smaller dimensions.

The data would normally be held in a spreadsheet program like Microsoft Office Excel or LibreOffice Calc, as in this example of an unselected table of multivariate statistical data from K.R. Gabriel in *Biometrika* 1971, **58**, 453–67.

Percent	Christian	Armenian	Jewish	Moslem	American	Shaafat	A-Tur	Silwan	Sur-Bahar
Toilet	98.2	97.2	97.3	96.9	97.6	94.4	90.2	94	70.5
Kitchen	78.8	81	65.6	73.3	91.4	88.7	82.2	84.2	55.1
Bath	14.4	17.6	6	9.6	56.2	69.5	31.8	19.5	10.7
Electricity	86.2	82.1	54.5	74.7	87.2	80.4	68.6	65.5	26.1
Water	32.9	30.3	21.1	26.9	80.1	74.3	46.3	36.2	9.8
Radio	73	70.4	53	60.5	81.2	78	67.9	64.8	57.1
TV set	4.6	6	1.5	3.4	12.7	23	5.6	2.7	1.3
Refrigerator	29.2	26.3	4.3	10.5	52.8	49.7	21.7	9.5	1.2

Such tables must be rectangular, with optional row and column labels, and all other cells filled with numerical data (missing data must be replaced by estimates). Often it would only be necessary to select cells containing numerical values from such a table by highlighting as follows.

Percent	Christian	Armenian	Jewish	Moslem	American	Shaafat	A-Tur	Silwan	Sur-Bahar
Toilet	98.2	97.2	97.3	96.9	97.6	94.4	90.2	94	70.5
Kitchen	78.8	81	65.6	73.3	91.4	88.7	82.2	84.2	55.1
Bath	14.4	17.6	6	9.6	56.2	69.5	31.8	19.5	10.7
Electricity	86.2	82.1	54.5	74.7	87.2	80.4	68.6	65.5	26.1
Water	32.9	30.3	21.1	26.9	80.1	74.3	46.3	36.2	9.8
Radio	73	70.4	53	60.5	81.2	78	67.9	64.8	57.1
TV set	4.6	6	1.5	3.4	12.7	23	5.6	2.7	1.3
Refrigerator	29.2	26.3	4.3	10.5	52.8	49.7	21.7	9.5	1.2

However, sometimes row and column labels could also be needed, when a labeled table with cells containing either labels or numerical values would be selected, as follows.

Percent	Christian	Armenian	Jewish	Moslem	American	Shaafat	A-Tur	Silwan	Sur-Bahar
Toilet	98.2	97.2	97.3	96.9	97.6	94.4	90.2	94	70.5
Kitchen	78.8	81	65.6	73.3	91.4	88.7	82.2	84.2	55.1
Bath	14.4	17.6	6	9.6	56.2	69.5	31.8	19.5	10.7
Electricity	86.2	82.1	54.5	74.7	87.2	80.4	68.6	65.5	26.1
Water	32.9	30.3	21.1	26.9	80.1	74.3	46.3	36.2	9.8
Radio	73	70.4	53	60.5	81.2	78	67.9	64.8	57.1
TV set	4.6	6	1.5	3.4	12.7	23	5.6	2.7	1.3
Refrigerator	29.2	26.3	4.3	10.5	52.8	49.7	21.7	9.5	1.2

Note that the dummy label in cell(1,1) is not used.

The structure of the default SIMFIT test file `houses.tf1` available after selecting [Statistics] from the SIMFIT main menu, followed by [Multivariate], then [Biplots] will now be explained.

First of all, note that SIMFIT is not constrained to work with spread sheet programs, and the data file format is more universal and much simpler, being a simple ASCII table of space separated numerical values with optional row and column labels. So the default test file `houses.tf1` contains the following table of observations.

98.2	97.2	97.3	96.9	97.6	94.4	90.2	94.0	70.5
78.8	81.0	65.6	73.3	91.4	88.7	82.2	84.2	55.1
14.4	17.6	6.0	9.6	56.2	69.5	31.8	19.5	10.7
86.2	82.1	54.5	74.7	87.2	80.4	68.6	65.5	26.1
32.9	30.3	21.1	26.9	80.1	74.3	46.3	36.2	9.8
73.0	70.4	53.0	60.5	81.2	78.0	67.9	64.8	57.1
4.6	6.0	1.5	3.4	12.7	23.0	5.6	2.7	1.3
29.2	26.3	4.3	10.5	52.8	49.7	21.7	9.5	1.2

Also, as the row and column labels would be required for a biplot, these are added to the test file as follows.

```
begin{labels}
Toilet
Kitchen
Bath
Electricity
Water
Radio
TV set
Refrigerator
Christian
Armenian
Jewish
Moslem
Am.Colony Sh.Jarah
Shaafat Bet-Hanina
A-Tur Isawyie
Silwan Abu-Tor
Sur-Bahar Bet-Safafa
end{labels}
```

An Excel macro called `simfit6.xls` is distributed with the SIMFIT package and it can output spreadsheet tables as correctly formatted SIMFIT data files from within Excel. Another easy way is to copy and paste the whole table directly into SIMFIT using the [Paste] option from the file selection control, or to copy and paste into program **maksim** which will then output a correctly formatted SIMFIT data file. However, if this course of action is to be followed, the following important restrictions may have to be noted.

1. There must be no missing values and every cell in the numeric part of the table must contain a valid number, except cell(1,1) which is ignored.

2. Data copied to the clipboard from a spreadsheet program will have tab separated columns and so SIMFIT will be able to perform numerous format conversion procedures interactively.

3. If spaces are used as column separators instead of tabs, the data must be in scientific format using full stops for decimal points not commas.

4. If spaces are used as column separators instead of tabs, there must no spaces in the labels, and any must be replaced by undercores before copying to the clipboard from a standard ASCII text editor such as **notepad**. For example, replace `time of day` by `time_of_day`, or `cycles per second` by `cycles_per_second`. This restriction does not matter with formatted SIMFIT data files as the row labels followed by the column labels are added as sequential lines between the begin{labels} and end{labels} section of the data file trailer.

The next figures illustrate typical biplots derived from `houses.tf1`.

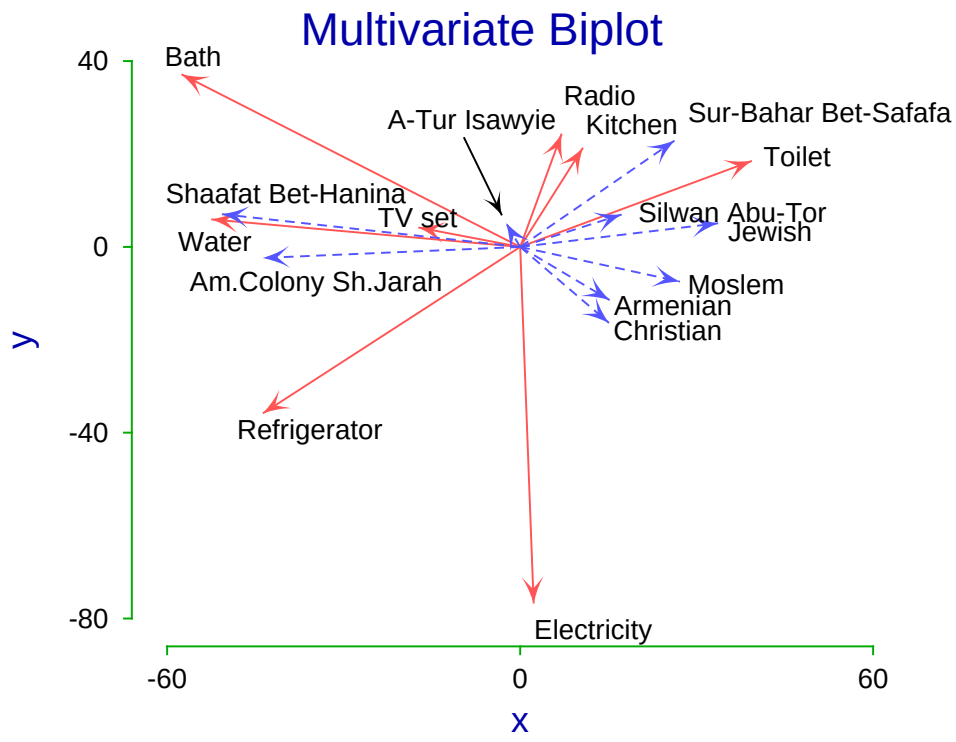

Figure 1: 2D Biplot

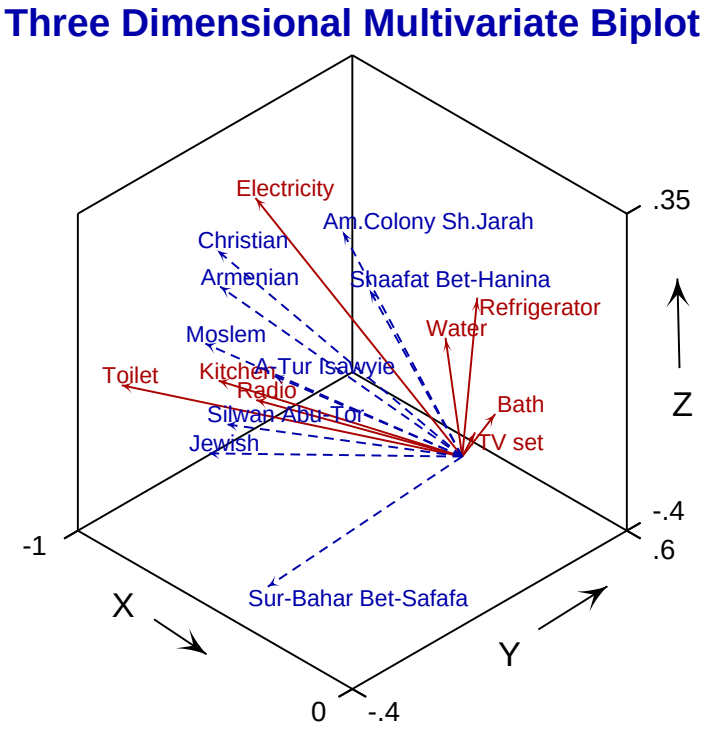

Figure 2: 3D Biplot

As with other projection techniques, such as principal components, it is necessary to justify that the number of singular values used to display a biplot does represent the data matrix adequately. To do this, consider the next table from the singular value decomposition of `houses.tf1`.

Proportion of total variance captured by singular values
Data file: `houses.tf1`, rank = 8

Index	σ_i	Fraction	Cumulative	σ_i^2	Fraction	Cumulative
1	499.393	0.7486	0.7486	249394	0.9631	0.9631
2	88.3480	0.1324	0.8811	7805.36	0.0301	0.9933
3	33.6666	0.0505	0.9315	1133.44	0.0044	0.9977
4	17.8107	0.0267	0.9582	317.222	0.0012	0.9989
5	12.8584	0.0193	0.9775	165.339	0.0006	0.9995
6	10.4756	0.0157	0.9932	109.738	0.0004	1.0000
7	3.37372	0.0051	0.9983	11.3820	0.0000	1.0000
8	1.15315	0.0017	1.0000	1.32974	0.0000	1.0000

In this example, it is clear that the first two or three singular values do represent the data adequately, and this is further reinforced by Figure 3 where the percentage variance represented by the successive singular values is plotted as a function of the singular value index. Here we see the cumulative variance $CV(i)$

$$CV(i) = \frac{100 \sum_{j=1}^{i} \sigma_j^2}{\sum_{j=1}^{k} \sigma_j^2}$$

plotted as a function of the index i, and such tables or plots should always be inspected to make sure that $CV(i)$ is greater than some minimum value (say 70 percent, for instance) for $i = 2$ or $i = 3$ as appropriate.

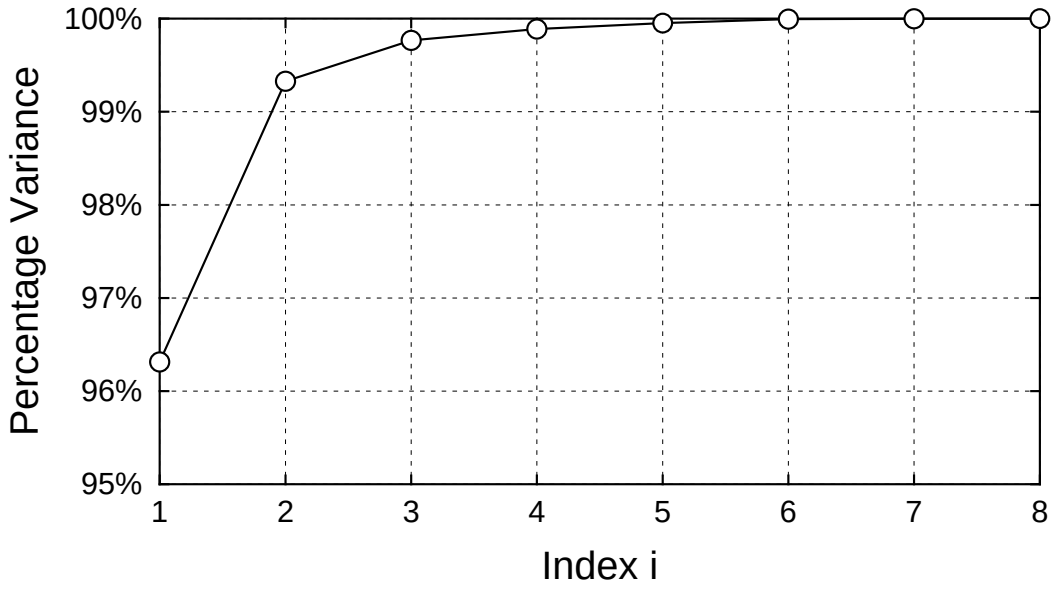

Figure 3: Cumulative Variance Plot

The theory behind the biplot options available in SimFIT will now be described.

Theory

The biplot is used to explore relationships between the rows and columns of any arbitrary matrix, by projecting the matrix onto a space of smaller dimensions using the singular value decomposition (SVD). It is based upon the fact that, as a n by m matrix X of rank k can be expressed as a sum of k rank 1 matrices as follows

$$X = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T,$$

then the best fit rank r matrix Y with $r < k$ which minimizes the objective function

$$\begin{aligned} S &= \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - y_{ij})^2 \\ &= \text{trace}[(X-Y)(X-Y)^T] \end{aligned}$$

is the sum of the first r of these rank 1 matrices. Further, such a least squares approximation results in the minimum value

$$S_{min} = \sigma_{r+1}^2 + \sigma_{r+2}^2 + \cdots + \sigma_k^2$$

so that the rank r least squares approximation Y accounts for a fraction

$$\frac{\sigma_1^2 + \cdots + \sigma_r^2}{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_k^2}$$

of the total variance, where k is less than or equal to the smaller of n and m, k is greater than or equal to r, and $\sigma_i = 0$ for $i > k$.

Figure 1 illustrates a biplot for the data in test file `houses.tf1`. The technique is based upon creating one of several possible rank-2 representations of of a n by m matrix X with rank k of at least two as follows. Let the SVD of X be

$$\begin{aligned} X &= U\Sigma V^T \\ &= \sum_{i=1}^{k} \sigma_i u_i v_i^T \end{aligned}$$

so that the best fit rank-2 matrix Y to the original matrix X will be

$$Y = \begin{pmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \\ \vdots & \vdots \\ u_{1n} & u_{2n} \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & \ldots & v_{1m} \\ v_{21} & v_{22} & \ldots & v_{2m} \end{pmatrix}.$$

Then Y can be written in several ways as GH^T, where G is a n by 2 matrix and H is a m by 2 matrix as follows.

1. General representation

$$Y = \begin{pmatrix} u_{11}\sqrt{\sigma_1} & u_{21}\sqrt{\sigma_2} \\ u_{12}\sqrt{\sigma_1} & u_{22}\sqrt{\sigma_2} \\ \vdots & \vdots \\ u_{1n}\sqrt{\sigma_1} & u_{2n}\sqrt{\sigma_2} \end{pmatrix} \begin{pmatrix} v_{11}\sqrt{\sigma_1} & v_{12}\sqrt{\sigma_1} & \ldots & v_{1m}\sqrt{\sigma_1} \\ v_{21}\sqrt{\sigma_2} & v_{22}\sqrt{\sigma_2} & \ldots & v_{2m}\sqrt{\sigma_2} \end{pmatrix}$$

2. Representation with row emphasis

$$Y = \begin{pmatrix} u_{11}\sigma_1 & u_{21}\sigma_2 \\ u_{12}\sigma_1 & u_{22}\sigma_2 \\ \vdots & \vdots \\ u_{1n}\sigma_1 & u_{2n}\sigma_2 \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & \ldots & v_{1m} \\ v_{21} & v_{22} & \ldots & v_{2m} \end{pmatrix}$$

3. Representation with column emphasis

$$Y = \begin{pmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \\ \vdots & \vdots \\ u_{1n} & u_{2n} \end{pmatrix} \begin{pmatrix} v_{11}\sigma_1 & v_{12}\sigma_1 & \dots & v_{1m}\sigma_1 \\ v_{21}\sigma_2 & v_{22}\sigma_2 & \dots & v_{2m}\sigma_2 \end{pmatrix}$$

4. User-defined representation

$$Y = \begin{pmatrix} u_{11}\sigma_1^{\alpha} & u_{21}\sigma_2^{\alpha} \\ u_{12}\sigma_1^{\alpha} & u_{22}\sigma_2^{\alpha} \\ \vdots & \vdots \\ u_{1n}\sigma_1^{\alpha} & u_{2n}\sigma_2^{\alpha} \end{pmatrix} \begin{pmatrix} v_{11}\sigma_1^{\beta} & v_{12}\sigma_1^{\beta} & \dots & v_{1m}\sigma_1^{\beta} \\ v_{21}\sigma_2^{\beta} & v_{22}\sigma_2^{\beta} & \dots & v_{2m}\sigma_2^{\beta} \end{pmatrix}$$

where $0 < \alpha < 1$, and $\beta = 1 - \alpha$.

To construct a biplot we take the n row effect vectors g_i and m column effect vectors h_j as vectors with origin at $(0, 0)$ and defined in the general representation as

$$\begin{aligned} g_i^T &= (u_{1i}\sqrt{\sigma_1}, u_{2i}\sqrt{\sigma_2}) \\ h_j^T &= (v_{1j}\sqrt{\sigma_1}, v_{2j}\sqrt{\sigma_2}) \end{aligned}$$

with obvious identities for the alternative row emphasis and column emphasis factorizations. The biplot consists of n vectors with end points at $(u_{1i}\sqrt{\sigma_1}, u_{2i}\sqrt{\sigma_2})$ and m vectors with end points at $(v_{1j}\sqrt{\sigma_1}, v_{2j}\sqrt{\sigma_2})$ so that interpretation of the biplot is then in terms of the inner products of vector pairs. That is, vectors with the same direction correspond to proportional rows or columns, while vectors approaching right angles indicate near orthogonality, or small contributions. Another possibility is to display a difference biplot in which a residual matrix R is first created by subtracting the best fit rank-1 matrix so that

$$\begin{aligned} R &= X - \sigma_1 u_1 v_1^T \\ &= \sum_{i=2}^{k} \sigma_i u_i v_i^T \end{aligned}$$

and this is analyzed, using appropriate vectors calculated with σ_2 and σ_3 of course. Again, the row vectors may dominate the column vectors or vice versa whatever representation is used and, to improve readability, additional scaling factors may need to be introduced. For instance, the previous figures used the residual matrix and scaling factors of -100 for rows and -1 for columns to reflect and stretch the vectors until comparable size was attained. To do this over-rides the default auto-scaling option, which is to scale each set of vectors so that the largest row and largest column vector are of unit length, whatever representation is chosen.

Biplots are most useful when the number of rows and columns is not too large, and when the rank-2 approximation is satisfactory as an approximation to the data or residual matrix. Note that biplot labels should be short, and they can be appended to the data file as with `houses.tf1`, or pasted into the plot as a table of label values. Fine tuning to re-position labels was necessary with these figures, and this can be done by editing the PostScript file in a text editor, or by using techniques described elsewhere for moving labels in scattergrams.

Sometimes, as with Figure 2, it is useful to inspect biplots in three dimensions. This has the advantage that three singular values can be used, but the plot may have to be viewed from several angles to get a good idea of which vectors of like type are approaching a parallel orientation (indicating proportionality of rows or columns) and which pairs of vectors i, j of opposite types are orthogonal (i.e., at right angles, indicating small contributions to x_{ij})