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

It is often necessary to fit models that are sums of sub–models, and some way of detecting the minimum number of sub–models required to explain the data is needed. Statistical tests like a sequential F test can help, but it is also useful to visualize the contribution of sub–models by plotting the sub–models at the same time as the best–fit curve. In SimFIT this technique is loosely referred to as graphical deconvolution. In short, given a model of the form

$$f(x,p) = \sum_{i=1}^{n} f_i(x,p)$$

where x is a vector of user–supplied independent variables and p is a vector of parameters to be estimated, then how should we attempt to calculate the contribution of sub–models f_i to the overall best–fit model, and thereby decide upon the minimum acceptable value for n, i.e., the minimum number of sub–models required.

As a typical example, consider the situation where a large sample is available and it is wished to fit a sequence of sums of normal distribution $cdfs$ as follows.

$$\begin{aligned} cdf_n(x) = {} & \frac{p_1}{p_{2n+1}\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left\{\frac{u-p_{n+1}}{p_{2n+1}}\right\}^2\right) du + \frac{p_2}{p_{2n+2}\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left\{\frac{u-p_{n+2}}{p_{2n+2}}\right\}^2\right) du + \cdots \\ & + \frac{p_n}{p_{3n}\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left\{\frac{u-p_{2n}}{p_{3n}}\right\}^2\right) du + p_{3n+1} \end{aligned}$$

However, particularly with machine–generated data, a histogram often has to be fitted using $pdfs$ in this form.

$$\begin{aligned} pdf_n(x) = {} & \frac{p_1}{p_{2n+1}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left\{\frac{x-p_{n+1}}{p_{2n+1}}\right\}^2\right) + \frac{p_2}{p_{2n+2}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left\{\frac{x-p_{n+2}}{p_{2n+2}}\right\}^2\right) + \cdots \\ & + \frac{p_n}{p_{3n}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left\{\frac{x-p_{2n}}{p_{3n}}\right\}^2\right) + p_{3n+1} \end{aligned}$$

For a sum of n such sub–models then up to $3n + 1$ parameters have to be estimated, namely

- $p_1, p_2, \ldots, p_n$ are positive partitioning fractions, which may be constrained, e.g., $\sum_{i=1}^{n} p_i = 1$
- $p_{n+1}, p_{n+2}, \ldots, p_{2n}$ are means of arbitrary sign,
- $p_{2n+1}, p_{2n+2}, \ldots, p_{3n}$ are positive standard deviations, and
- p_{3n+1} is an arbitrary background correction factor that is sometimes required.

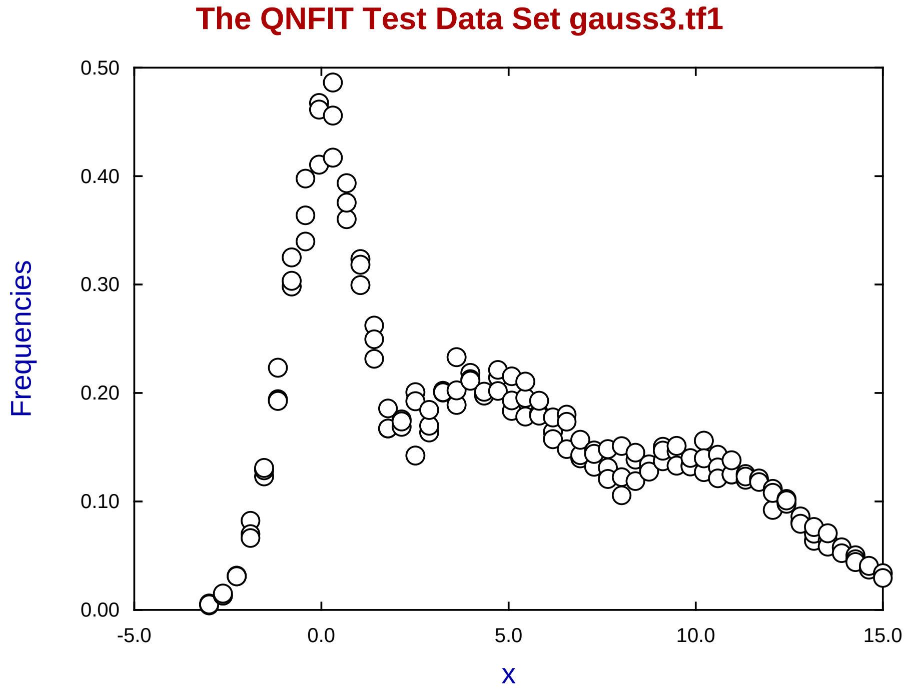
Several points should be noted about fitting this model.

1. Fitting the cdf uses a data set that is unique up to order, but the shape is somewhat featureless making it difficult to guess parameter starting estimates and limits.

2. Fitting the pdf is not unique as it depends on the number of histogram bins but, when the peaks are well–separated, it is easier to guess parameters starting estimates and assess goodness of fit visually.

3. The numbering of sub–models is arbitrary as regards permutations, so the starting estimates and parameter limits must be chosen with considerable care to limit the search directions for minimizing the objective function in order to avoid ambiguity. This consideration also applies when fitting similar models like sums of saturation functions or exponentials.

However, models such as these are not always used for statistical analysis of a mixed sample but rather as simple empirical models in an attempt to resolve the contribution of individual signals to an overall profile. So, in order to illustrate this procedure in such an application, we shall consider the data set contained in the SimFIT test file `gauss3.tf1` which contains triplicates as illustrated next.



From inspecting this profile we see that there are at least three distributions involved. These appear to make similar contributions so we could guess that

p_1, p_2 and p_3 would be of order unity, while
$0 \approx p_4 << p_5 << p_6 << 12$,
$1 \approx p_7 << p_8 << p_9 << 5$, and it would be safe to fix the constant term so that
$p_{10} = 0$.

So attached to the end of `gauss3.tf1` will be found these limits and starting estimates, allowing this data set to be fitted in EXPERT mode where such estimates and limits are read from the data file, and which greatly facilitates fitting by SimFIT program **qnfit**.

```
begin{limits}
  0,   0.5,   2
  0,   1.5,   2
  0,   0.5,   2
 -2,   0.0,   2
  2,   4.0,   6
  8,  10.0,  12
0.1,   1.0,   2
  1,   2.0,   3
  2,   3.0,   4
  0,   0.0,   0
end{limits}
```

To appreciate how such an analysis would be conducted proceed as follows.

How to plot a graphical deconvolution

1. Open SimFIT using either **w_simfit.exe** for the 32-bit version, or **x64_simfit.exe** for the 64-bit version.
2. From the main SimFIT menu press the [A/Z] option.
3. Scroll down the list displayed and open program **qnfit**.
4. Accept the default options and select the option to fit one function of one variable.
5. Read in the test file called `gauss3.tf1` by pressing the [Demo] button on the file opening dialogue and scrolling down the list provided.
6. Choose the option to fit Gaussian $pdfs$ (spikes) with no constant term.
7. When asked how many terms are required input a 3, i.e., choose to fit a sum of three terms.
8. Select the option to run in the EXPERT mode which reads starting estimates and limits off the data file `gauss3.tf1`.
9. Proceed to fitting.

A summary of the optimization procedure and preliminary comments about the goodness of fit are given and then the parameter estimates from fitting are displayed and output as text to the results file as below.

```
 Best-fit parameters for curve-fit   1 using LBFGSB

No.   Low-Limit High-Limit    Value        Std.Error    Lower95%cl    Upper95%cl    p
  1   0.000E+00   2.000E+00  9.07541E-01  2.16240E-02  8.64792E-01  9.50291E-01  0.0000
  2   0.000E+00   2.000E+00  1.16433E+00  4.21732E-02  1.08096E+00  1.24770E+00  0.0000
  3   0.000E+00   2.000E+00  9.25185E-01  3.01303E-02  8.65619E-01  9.84750E-01  0.0000
  4  -2.000E+00   2.000E+00 -7.29763E-02  1.55718E-02 -1.03761E-01 -4.21918E-02  0.0000
  5   2.000E+00   6.000E+00  3.74510E+00  5.08157E-02  3.64464E+00  3.84556E+00  0.0000
  6   8.000E+00   1.200E+01  1.02774E+01  9.64127E-02  1.00868E+01  1.04680E+01  0.0000
  7   1.000E-01   2.000E+00  9.26404E-01  1.43311E-02  8.98073E-01  9.54736E-01  0.0000
  8   1.000E+00   3.000E+00  2.34330E+00  7.05668E-02  2.20380E+00  2.48281E+00  0.0000
  9   2.000E+00   4.000E+00  2.76906E+00  6.26372E-02  2.64523E+00  2.89289E+00  0.0000
      parameter(10) is the excluded constant term
 For 50,90,95,99% con. lim. using [parameter value +/- t(alpha/2)*std.err.]
 t(.25) = 0.676, t(.05) = 1.656, t(.025) = 1.977, t(.005) = 2.611
```

Later, after using the [Results] option from the main SimFIT menu, these results can be extracted as in the following table for inclusion in documents.

Best-fit parameters for curve-fit 1 using LBFGSB

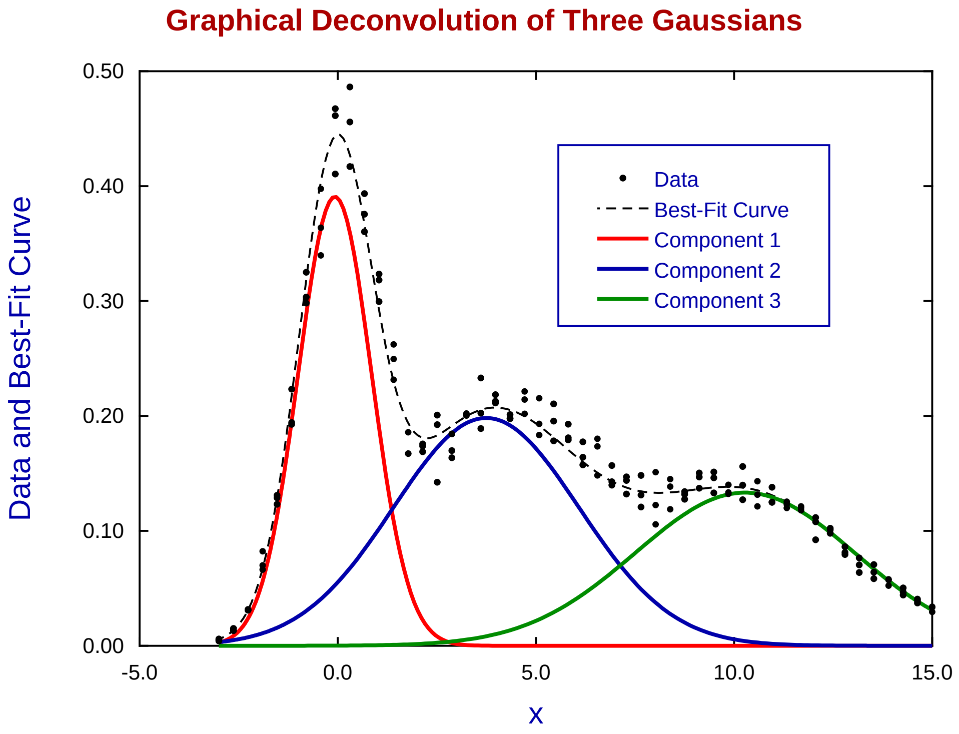
Number	Low-Limit	High-Limit	Value	Std.Error	Lower95%cl	Upper95%cl	p
1	0.0	2.0	0.90754	0.021624	0.86479	0.95029	0.0000
2	0.0	2.0	1.16433	0.042173	1.08096	1.24770	0.0000
3	0.0	2.0	0.92519	0.030130	0.86562	0.98475	0.0000
4	-2.0	2.0	-0.07298	0.015572	-0.10376	-0.04219	0.0000
5	2.0	6.0	3.74510	0.050816	3.64464	3.84556	0.0000
6	8.0	12	10.2774	0.096413	10.0868	10.4680	0.0000
7	0.1	2.0	0.92640	0.014331	0.89807	0.95474	0.0000
8	1.0	3.0	2.34330	0.070567	2.20380	2.48281	0.0000
9	2.0	4.0	2.76906	0.062637	2.64523	2.89289	0.0000

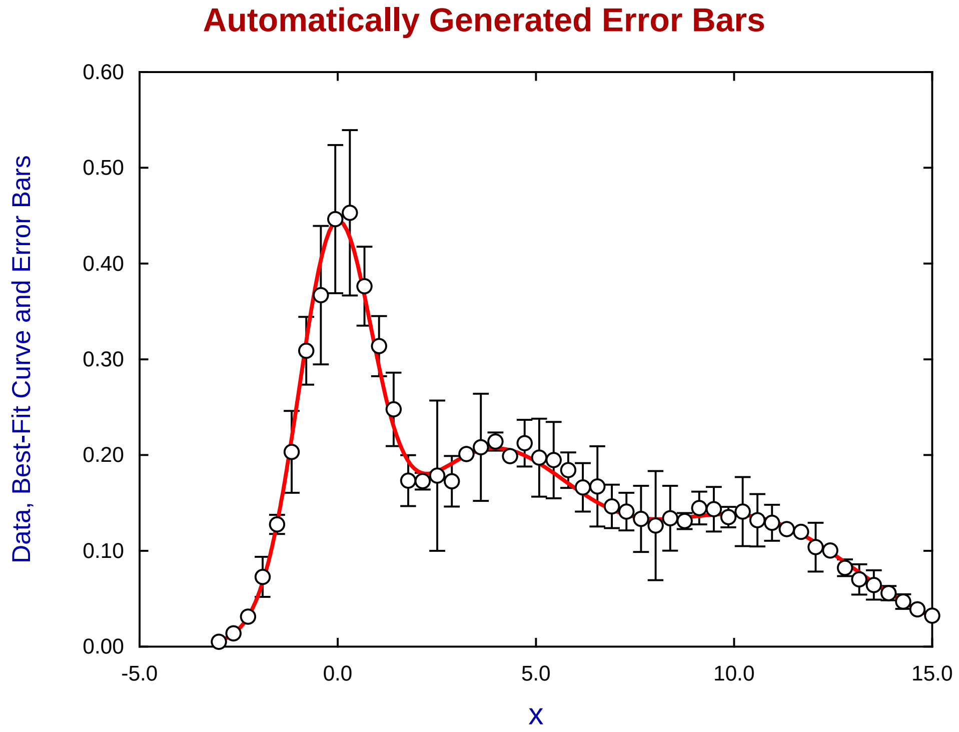
parameter(10) is the excluded constant term

For 50,90,95,99% con. lim. using [parameter value +/- t(α/2)*std.err.]

t(.25) = 0.676, t(.05) = 1.656, t(.025) = 1.977, t(.005) = 2.611

After proceeding through subsequent menus providing options for graphical display and goodness of fit analysis a final end–of–analysis menu is reached. From this the option to view the graphical deconvolution can be selected to create the next graph showing the data, best–fit curve, and individually contributing components, or the subsequent plot illustrating error bars.


Graphical Deconvolution of Three Gaussians
Data and Best-Fit Curve
x
Data
Best-Fit Curve
Component 1
Component 2
Component 3


Automatically Generated Error Bars
Data, Best-Fit Curve and Error Bars
x