Non-linear regression. The Gauss-Newton method

Non-linear methods are challenging. It is often difficult to obtain appropriate convergence using traditional methods. The principles of H-methods have been applied to many areas within non-linear modelling. It has been found that these are superior to traditional methods. Here we shall consider estimation of parameters in non-linear models.

In non-linear regression the parameters of the model appear in a non-linear way. The models are often derived from some theoretical considerations. Examples of how models are derived are as follows.

Solution to differential equations:

$$y(x) = \theta_1/(\theta_1 - \theta_2)\{\exp(-\theta_2 x) - \exp(-\theta_1 x)\}$$

$$y(x) = \alpha/\{1 + \exp[-(\lambda + kx)/\delta]\}^{\delta}$$

Models from approximations:

$$y(x) = \beta_0 + \beta_1 \exp(\beta_2 x)$$

Models from the shape of data:

$$y(x) = \beta_0 + \beta_1 \exp(\beta_2 (x-\mu_1)) + \beta_3 \exp(\beta_4 (x-\mu_2))$$

In these models there is only one x-variable. When there are several x-variables and many parameters, it may be a difficult task to estimate the parameters from the given data. A standard procedure is to use the Gauss-Newton procedure, which is briefly described in the following.

Suppose that the mathematical model is $y=f(\mathbf{x};\boldsymbol{\beta})$. The parameters $\boldsymbol{\beta}$ are found by minimizing the sum of squared errors,

$$\text{minimize } (\mathbf{y} - f(\mathbf{x};\boldsymbol{\beta}))^T(\mathbf{y} - f(\mathbf{x};\boldsymbol{\beta}) .$$

If we differentiate the expression and equate the derivative to zero, we get

$$-2F(\mathbf{x};\boldsymbol{\beta})^T\mathbf{y} + 2F(\mathbf{x};\boldsymbol{\beta})^T\mathbf{f} = \mathbf{0}$$

where $F(\mathbf{x};\boldsymbol{\beta})=\partial f(\mathbf{x};\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. This leads to the equations

(13) $$\mathbf{F}(\mathbf{x};\boldsymbol{\beta})^T\mathbf{f} = \mathbf{F}(\mathbf{x};\boldsymbol{\beta})^T\mathbf{y}.$$

If the model is linear, $\mathbf{F}$ is the design matrix $\mathbf{X}$, $\mathbf{f}$ is $\mathbf{X}\boldsymbol{\beta}$ and the normal equations to be solved are $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}=\mathbf{X}^T\mathbf{y}$. When $\mathbf{f}$ is non-linear, we expand it by a Taylor series,

(14) $$\mathbf{f}(\mathbf{x};\boldsymbol{\beta}) \cong \mathbf{f}(\mathbf{x};\boldsymbol{\beta}_0) + \mathbf{F}(\mathbf{x};\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$$

If we insert the right hand side of (8) into (7) and rearrange terms, we get

(15) $$\mathbf{F}(\mathbf{x};\boldsymbol{\beta}_0)^T\mathbf{F}(\mathbf{x};\boldsymbol{\beta}_0)\,(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = \mathbf{F}(\mathbf{x};\boldsymbol{\beta}_0)^T(\mathbf{y} - \mathbf{f}(\mathbf{x};\boldsymbol{\beta}_0)).$$

This has the form of the normal equations with $\mathbf{X}=\mathbf{F}(\mathbf{x};\boldsymbol{\beta}_0)$. These equations are solved for $\Delta\boldsymbol{\beta}=\boldsymbol{\beta} - \boldsymbol{\beta}_0$. The new solution is then $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \Delta\boldsymbol{\beta}$. The equations (14) and (15) are iterated with $\boldsymbol{\beta}_0$ in (14) is now the revised solution from (15). The iterations stop, when the $\boldsymbol{\beta}$ does not change or the residual sum of squares does not reduce. When there are many parameters, it may be difficult to obtain convergence, if the linear least squares solution is used. There have been suggested many methods to obtain convergence. One approach, the Hartley corrections, is not to use all of $\Delta\boldsymbol{\beta}$ as a correction, but $\Delta\boldsymbol{\beta}/2$ or $\Delta\boldsymbol{\beta}/2^k$, k=0,1,... and see if the residual sum of squares gets diminished. Another approach is to use Ridge regression (also called Marquardt) corrections, where we start by letting the Ridge constant be e.g., $\lambda= 10^{-8}$. If the residual sum of squares does not diminish, we increase λ, e.g., 10λ. If we can not find λ that reduces the residual sum of squares, the iterations stop. There are many other methods that have been proposed.

The H-method computes the solution vector to (15) for each dimension,

$$\mathbf{b}_a= d_1\, \mathbf{r}_1\, q_1 + \ldots + d_a\, \mathbf{r}_a\, q_a, \qquad \text{for a=1,...,A (or K).}$$

Then a trial for a solution is computed as

$$\Delta\boldsymbol{\beta} = \text{step} \times \mathbf{b}_a, \qquad \text{for step}=1/2^5,\ 1/2^4,\ldots,1/2^0=1.$$

For each of A×6 solution vectors, $\Delta\boldsymbol{\beta}$'s, the residual sum of squares, $|\mathbf{y} - f(\mathbf{x};\boldsymbol{\beta})|^2$, is computed. The dimension, $a=a_0$, and $\text{step}=\text{step}_0$, are chosen that give the minimum value of the residual sum of squares. When $\Delta\boldsymbol{\beta}$ has been found, the iteration starts over again as described above. If none of the A×6 solution vectors can decrease the residual sum of squares, the computations stop.

At convergence we have score and loading vectors from the last iteration. This can be used for studying the properties of the solution vector.

We shall consider an example, where for most methods it has been difficult to obtain appropriate solutions at convergence.

Sum of Gaussian curves. In analytical chemistry we sometimes see need for estimating the parameters in a model that is a sum of Gaussian curves. An example of such a function is the model,

$$(16) \qquad y(x) = c_1 \exp(-(x - a_1)^2/b_1) + c_2 \exp(-(x - a_2)^2/b_2) + c_3 \exp(-(x - a_3)^2/b_3).$$

The matrix **F** of the differential of the parameters will have 9 columns. If the estimates of the parameters (a_i,b_i,c_i) are not very close to the true values, the matrix **F** will be close to singular. In these cases traditional algorithms will not converge or give bad estimates. If we on the other hand use the H-method at each step, we typically get convergence. Let us consider a numerical example. Consider the data that are shown in Figure 15. The figure shows the experimental data. The scatter of points clearly indicates three peaks. It is therefore natural to expect the model (16) to be appropriate. An initial estimate of the parameters is

$$(a_1,b_1,c_1,\ a_2,b_2,c_2,\ a_3,b_3,c_3) = (4.4,\ 0.11,\ 0.6,\ 4.8,\ 0.30,\ 0.7,\ 5.8,\ 0.20,\ 1.1)$$

The initial values are found by identifying each Gaussian curve approximately with the peaks in data. With these initial values the matrix **F** is not close to be singular. But the condition number, the ratio of the largest and smallest singular value of **F**, is 261.8. This value of the condition number is sufficiently large that it gives an unstable solution for many methods.

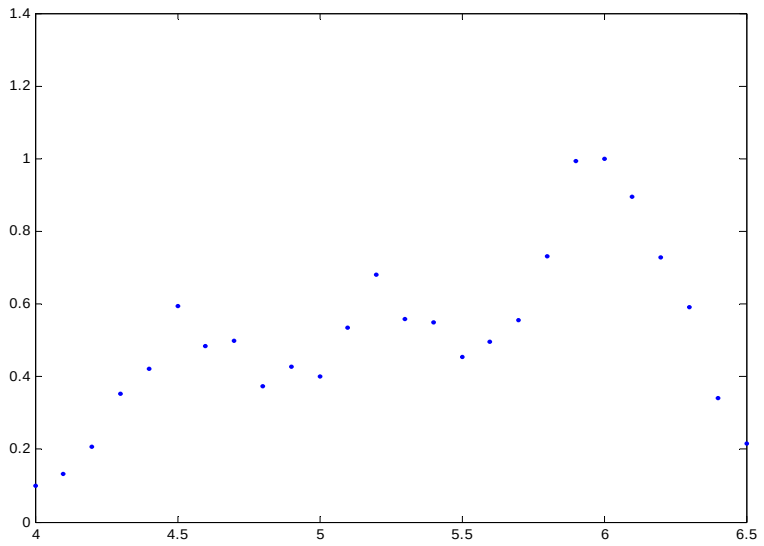

Figure 15. Experimental data, where sum of three Gaussian curves are expected to fit data.

When we use the exact solution and modification of it, we get after 100 iterations the results shown in figure 16.

The parameter estimates are as follows.

$(a_1,b_1,c_1, a_2,b_2,c_2, a_3,b_3,c_3)$=(2.65 0.01 -0.03 5.09 1.12 0.54 6.05 0.10 0.75)

The negative value c_1=-0.03 indicates that we are subtracting the first Gaussian curve from the other two. This is not the interpretation that can be expected, when looking at Figure 16. The residual sum of squares is $\Sigma(y_i - \hat{y}_i)^2$=0.160. This shows that we have obtained a relatively good fit, although the shape of the estimated function is not as expected.

In Figure 17 we show the results of applying the H-method. At each step there is computed 9×6=54 possible solutions,

$\Delta\boldsymbol{\beta} = \text{step} \times \mathbf{b}_a$,

29 iterations are needed to get convergence. At most iterations the dimension chosen is 3-5, and step-size 1, ½ or ¼. The parameter estimates are now as follows.

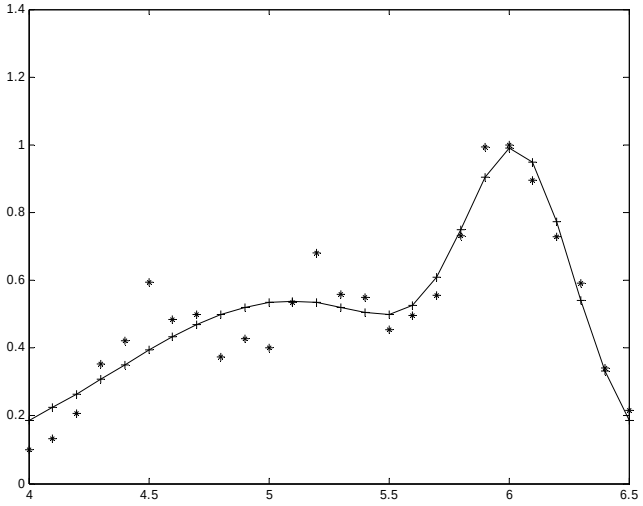

Figure 16. Estimated curve using exact solutions and modifications of them.

$(a_1,b_1,c_1, a_2,b_2,c_2, a_3,b_3,c_3)$=(4.55 0.14 0.54 5.23 0.09 0.59 6.00 0.15 0.99)

These values look natural. The interpretation of e.g., the first set is that the mean value of the first curve is at 4.55, the standard deviation is σ=0.26 ($\sigma^2=b_1/2$) and the first curve has the weight c=0.54. The fit is here much better, $\Sigma(y_i - \hat{y}_i)^2$=0.034.

In conclusion, the H-method has been applied to many non-linear modelling tasks. The general results are that the H-method is superior to full-rank solution methods, which are standard in program packages and textbooks on non-linear modelling. When the H-method is applied, the convergence results are superior, the interpretation the parameter is better and predictions are more stable than obtained by traditional methods of non-linear modelling.

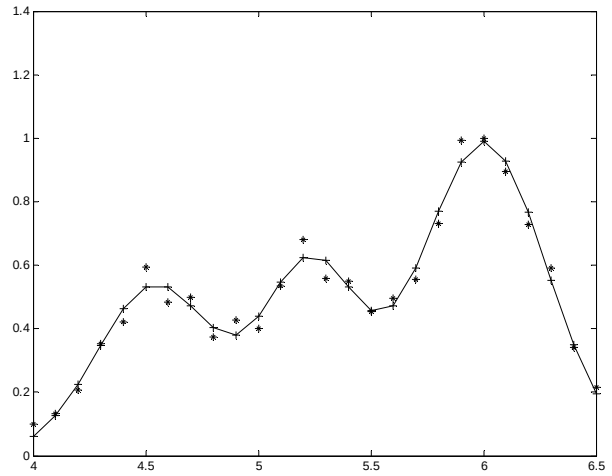

Figure 17. Estimated curve using low rank solutions of the H-method.