SOME EXPERIENCES IN THE ESTIMATION OF PARAMETERS IN NON-LINEAR DIFFERENTIAL EQUATIONS

J. G. P. BARNES
Reading, England

The author describes a procedure developed by himself and his colleagues for obtaining estimates of the parameters of rate equations, together with information about confidence regions for the estimates. The program has been used successfully for processing results from the chemical engineering industry, with highly non-linear model systems, particularly since temperature was a variable, and the "rate constants" were non-linear combinations of other constants. In biochemical situations, in which investigations are almost always at constant temperature, the non-linearity should not be so extreme, and the procedure may well be capable of dealing with more than 5 to 7 parameters for which it is recommended.

1. Introduction

This note describes in general terms some experiences with the estimation of parameters in non-linear differential equations describing chemical reactions. The original programs described were written in Mercury Autocode with P1G sequences, and run on the ICI Mercury computer in 1962–63. The general work has never been published externally but some of the problems "solved" have been published individually.

2. The problem

The exact chemical systems in which we were interested need not concern us, but in general terms they involved of the order of 5 chemicals interacting in about as many reactions. The equations were very non-linear; some of the species appeared in the equations as terms of non-integral order. In most cases the unknown parameters were the "rate constants", but these were often expressed as $\alpha \exp(-\beta/RT)$, with α and β treated as separate unknowns. For this latter to be possible it was of course essential to have data at different temperatures. Occasionally an unknown parameter was the order of a species in a reaction. Thus not only were the differential equations non-linear, but the unknown parameters themselves entered into the equations in a non-linear manner.

In a typical situation data would be available at, say, 6 different temperatures, with perhaps some repetition with different initial concentrations, giving about a dozen sets in all. Each set would consist of measurements of some (about half) of the reactants at perhaps 10 values of time. Thus we would have of the order of 300 measurements from which to estimate 5–7 unknown parameters. In addition to the unknown parameters, we desired some information on confidence regions.

3. Integration

The equations were solved by the Kutta-Merson explicit method with automatic step-length control. Experience in step-length control led one to be cautious about increasing the step length until the algorithm had indicated for several steps in succession that an increase was feasible. Apart from the above proviso no difficulty was experienced in solving the equations, since they were not too stiff.

4. Parameter estimation

In the early stages the method due to Rosenbrock [1] was used. This proved reliable but tedious. Of the

order of 200 integrations were required for each problem and this took about 6 hours on Mercury (add time 60 μsec). The main disadvantage of the method was its first order convergence. Thus each extra decimal digit in the results cost the same time and consequently it was very difficult to know when to stop. No feeling of finality was obtained. The other disadvantage was the lack of any naturally-occurring information on confidence regions.

In order to combat the above shortcomings the method described in the specification below was evolved. This proved most satisfactory; it gave second order convergence, and confidence regions. Its only disadvantage was its sensitivity to initial guesses for the unknown parameters. As a compromise the Rosenbrock method was used to obtain a good starting point for the second order method.

Although approximate confidence regions were obtained they were often very difficult to interpret. All too often one or more linear combinations of unknowns were not determined at all. The only way round this is, of course, the design of adequate experiments in the first place.

5. Conclusions

The methods to be described proved satisfactory provided the determination of too many parameters was not attempted. We usually got into trouble if more than 7 were involved; 10 would be quite hopeless. 5 were almost always easy.

6. Epilogue

The current situation (September 1968) regarding the programs is as follows.

i) The original Mercury programs are now destroyed.

ii) A large program in K-Autocode for the English Electric (ICL) KDF9 computer exists. This program incorporates both methods, and further information is available from the author.

iii) The second method has been preserved for posterity in the form of an Algol procedure. A detailed specification follows, and further information may be obtained from the author of this note,

FEBS are very grateful to Dr. Barnes and Dr. J.L. Wales, who developed the procedure, and to ICI Central Instruments Research Division, for permission to publish the specification which follows.

SPECIFICATION OF ALGOL PROGRAM FOR NON-LINEAR OPTIMISATION

1. Purpose

Given the functions (in general non-linear):

$$\epsilon_\iota = \epsilon_\iota(u_1, u_2, ..., u_N) \qquad [\iota = 1, ..., U; U > N] ,$$

the procedure minimises the function

$$E(u_1, u_2, ..., u_N) = \sum_{\iota=1}^{U} \epsilon_\iota^2 = \boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}$$

with respect to the unknown parameters u_j. The functions may be defined implicitly, it being necessary only to provide a procedure "eps" which evaluates the ϵ_ι for given values of the u_j.
The procedure is primarily intended for use in the case $U \gg N$.

Procedure heading

```
procedure nonlinest (N,U,Fa,b,conv,dumax,its,start,Print,
          u,ul,uu,L,Q,eps,dev,f1,f2,f3,f4);
value     N,U,Fa,b,conv,dumax,its,start,Print,dev,f1,f2,
          f3,f4;
real      Fa,b,conv,dumax;
integer   N,U,its,Print,f1,f2,f3,f4,dev;
```

```
array     u,ul,uu,L,Q;
boolean   start;
procedure eps;
```

2. Description

Suppose that $\mathbf{u}^o$ is a trial vector (base point) and that $\delta\mathbf{u}$ is the required correction.
The residual functions ϵ_i are approximated by a linear function of the parameters

$$\epsilon_i(\mathbf{u}^o + \delta\mathbf{u}) = \epsilon_i(\mathbf{u}^o) + \sum_{j=1}^{N} a_{ij}\,\delta u_j \qquad (1)$$

then

$$E = \Sigma\epsilon_i^2(\mathbf{u}^o + \delta\mathbf{u})$$

$$= \Sigma\epsilon_i^2(\mathbf{u}^o) - 2\mathbf{a}'\mathbf{b}\delta\mathbf{u} + \delta\mathbf{u}'\mathbf{a}'\mathbf{a}\delta\mathbf{u} \qquad (2)$$

where $\mathbf{a} = [a_{ij}]$ a $U \times N$ matrix
$\mathbf{b} = [-\epsilon_i(u^o)]$ a U vector

This approximating function to E has a minimum at the point given by the normal equations

$$\mathbf{c}\delta\mathbf{u} = \mathbf{d} \qquad (3)$$

where

$$\mathbf{c} = \mathbf{a}'\mathbf{a}, \qquad \mathbf{d} = \mathbf{a}'\mathbf{b} \qquad (4)$$

Hence

$$\delta\mathbf{u} = \mathbf{c}^{-1}\mathbf{d} \qquad (5)$$

On the linear theory $\mathbf{u}^o + \delta\mathbf{u}$ so determined would be the required solution and the minimum value of E attained there would be

$$E_1 = E_0 - \mathbf{d}'\mathbf{c}^{-1}\mathbf{d} \qquad (6)$$

(E_1 is printed out for comparison with the actual value of $E(\mathbf{u}^o + \delta\mathbf{u})$).
In general, the point $\mathbf{u} + \delta\mathbf{u}$ is not the required point and a simple search is then performed in an attempt to find a value of m such that $\mathbf{u} = \mathbf{u}^o + m\delta\mathbf{u}$ is a better solution of the non-linear problem. The whole process is then repeated using this point $\mathbf{u}$ as the base point $\mathbf{u}^o$ for the next iteration.
The matrix elements a_{ij} are evaluated by perturbing the parameters $\mathbf{u}$ and using a central difference formula. However, for reasons which will become apparent later, the perturbations are not in fact along the axes $\{u\}$ but along a set of axes $\{v\}$ which in the final stages are the eigenvectors of the quadratic approximation to the response surface $E(\mathbf{u})$.
So, consider a linear transformation

$$\delta\mathbf{v} = \mathbf{P}\delta\mathbf{u}$$

where

$$\mathbf{P} = \mathbf{L}\mathbf{Q}'$$

and $\mathbf{L}$ is a diagonal matrix
$\mathbf{Q}$ is an orthogonal matrix
so that

$$\mathbf{P}^{-1} = \mathbf{Q}\mathbf{L}^{-1}$$

Equation (1) may then be written

$$\epsilon_i = \epsilon_i^o + \mathbf{a}\mathbf{P}^{-1}\delta\mathbf{v} = \epsilon_i^o + \mathbf{A}\delta\mathbf{v} \qquad (7)$$

where we use capital letters to denote matrices in the $\{v\}$ co-ordinate system.
The matrix $\mathbf{A}$ is evaluated by perturbations along the $\{v\}$ axes. The magnitude $\delta\mathbf{v}$ of the perturbations is discussed below.

We use

$$A_{ij} = \{\epsilon_i(\mathbf{v}^o + \delta\mathbf{v}_j) - \epsilon_i(\mathbf{v}^o - \delta\mathbf{v}_j)\}/2\delta v \qquad (8)$$

where $\delta\mathbf{v}_j$ is the vector whose only non-zero element is δv in the jth entry.
The perturbations are actually evaluated in the $\{u\}$ system and the corresponding perturbations $\delta\mathbf{u}_j$ are given by

$$\delta\mathbf{u}_j = \mathbf{Q}\mathbf{L}^{-1}\delta\mathbf{v}_j$$

and form the columns of the perturbation matrix $\Delta\mathbf{U}$.

We then evaluate

$$\mathbf{C} = \mathbf{A}'\mathbf{A} = (\mathbf{P}^{-1})'\mathbf{a}'\mathbf{a}\mathbf{P}^{-1} = (\mathbf{P}^{-1})'\mathbf{c}\mathbf{P}^{-1} \qquad (9)$$

$$\mathbf{D} = \mathbf{A}'\mathbf{b} = (\mathbf{P}^{-1})'\mathbf{a}'\mathbf{b} \quad = (\mathbf{P}^{-1})'\mathbf{d} \qquad (10)$$

and

$$\delta \mathbf{v} = \mathbf{C}^{-1}\mathbf{D}$$

and finally

$$\delta \mathbf{u} = \mathbf{P}^{-1}\delta \mathbf{v} = \mathbf{Q}\mathbf{L}^{-1}\delta \mathbf{v} \tag{12}$$

We also need $\mathbf{c}$ in order to evaluate its eigensystem

$$\begin{aligned}\mathbf{c} &= \mathbf{P}'\mathbf{C}\mathbf{P} \\ &= \mathbf{Q}\mathbf{L}\mathbf{C}\mathbf{L}\mathbf{Q}'\end{aligned} \tag{13}$$

and $\mathbf{d}'\mathbf{c}^{-1}\mathbf{d} = \mathbf{D}'\mathbf{C}^{-1}\mathbf{D}$ in order to estimate the minimum value E_1 of E from equation (6).
At each iteration the eigenvalues and eigenvectors of the matrix $\mathbf{c}$ are evaluated and provide the matrices $\mathbf{L}$ and $\mathbf{Q}$ of the transformation for the next iteration. $\mathbf{L}$ is taken to be the diagonal matrix whose diagonal elements are

$$\lambda_1^{\frac{1}{2}}, \lambda_2^{\frac{1}{2}}, \lambda_3^{\frac{1}{2}}, \ldots, \lambda_N^{\frac{1}{2}}$$

and $\mathbf{Q}$ is the matrix whose ith column is the (normalised) eigenvector corresponding to the eigenvalue λ_i.
Suppose that the matrix $\mathbf{c}$ is unchanged from one stage to the next.
We can then express $\mathbf{c}$ as

$$\mathbf{c} = \mathbf{Q}\mathbf{L}^2\mathbf{Q}' \tag{14}$$

and so

$$\mathbf{C} = \mathbf{P}\mathbf{c}\mathbf{P}^{-1} = \mathbf{L}^{-1}\mathbf{Q}'\mathbf{Q}\mathbf{L}^2\mathbf{Q}'\mathbf{Q}\mathbf{L}^{-1} = \mathbf{1} \tag{15}$$

In general the matrix $\mathbf{C}$ tends to the unit matrix at later stages in the process.
One object of the transformation is now clear. If the response surface $E = E(\mathbf{u})$ is ill-conditioned, the matrix $\mathbf{c}$ will be so also (i.e., the spread of its eigenvalues will be large) and the calculation of $\mathbf{c}^{-1}$ will give spurious results. The matrix $\mathbf{C}$, however, is always well-conditioned and the evaluation of $\mathbf{C}^{-1}$ presents no problem.
The second object of the transformation concerns the statistics of the results. It can be shown that under certain assumptions, the confidence region at level α is the ellipsoidal region

$$(\mathbf{u}-\hat{\mathbf{u}})'\mathbf{c}(\mathbf{u}-\hat{\mathbf{u}}) \leqslant \left(\frac{N}{U-N}\right)\hat{E}F_\alpha(N, U-N) \tag{16}$$

where $\hat{\mathbf{u}}$ is the final estimate of $\mathbf{u}$ and $\hat{E} = E(\hat{\mathbf{u}})$.
$F_\alpha(N, U-N)$ is the α-point of the F distribution with N and $U-N$ degrees of freedom.
The size of perturbations used for the evaluation of $\mathbf{A}$ at each stage are such that the points at which E is evaluated all lie on the ellipsoid

$$(\mathbf{u}-\mathbf{u}_0)'\mathbf{c}(\mathbf{u}-\mathbf{u}_0) = \left(\frac{N}{U-N}\right)E_0F_\alpha(N, U-N) \tag{17}$$

and so, in the final stage, become the end points of the principal axes of the confidence ellipsoid.
In the $\{v\}$ co-ordinate system equation (17) becomes simply (by putting $\mathbf{C} = \mathbf{I}$ from equation (15)):

$$(\mathbf{v}-\mathbf{v}_0)'(\mathbf{v}-\mathbf{v}_0) = \left(\frac{N}{U-N}\right)E_0F_\alpha(N, U-N) \tag{18}$$

showing that the magnitude of the required perturbations in the $\{v\}$ system is

$$|\mathbf{dv}| = \sqrt{\left(\frac{N}{U-N}\right)E_0F_\alpha(N, U-N)} \tag{19}$$

Simple constraints of the form

$$ul_i \leqslant u_i \leqslant uu_i \tag{20}$$

are applied (and if not needed should be set very wide).
The procedure nonlinest does not, however, find the point having the minimum value of E in the constrained region if the absolute minimum is outside that region. That is, nonlinest does not solve the corresponding constrained problem. The constraints are intended to prevent nonlinest from trying values of $\mathbf{u}$ such that $\boldsymbol{\varepsilon}(\mathbf{u})$ might be indeterminate. (For example, the evaluation of $\boldsymbol{\varepsilon}$ might involve integrations which do not converge for certain values of $\mathbf{u}$).
If the estimate $\mathbf{u} = \mathbf{u}^0 + \delta\mathbf{u}$ violates the constraints, nonlinest prints the letter "c" (it may print several, the number depending upon which violation is detected first), and tries instead the point $\mathbf{u} = \mathbf{u} + K\delta\mathbf{u}$ where

K is chosen so that $\mathbf{u}$ is just inside the allowed region ($K = 0.999$ of the value which would result in $\mathbf{u}^o + K\delta\mathbf{u}$ being on the boundary). So the first trial value $\delta\mathbf{u}^{T_1}$ is $K\delta\mathbf{u}$ or $\delta\mathbf{u}$ according as the constraints were or were not violated. A simple search for a better point is now carried out along the line joining the points $\mathbf{u}^o$ and $\mathbf{u}^o + \delta\mathbf{u}$.

This search involves trying the step

$$\delta\mathbf{u}^T = m\,\delta\mathbf{u}^B$$

where $\delta\mathbf{u}^B$ is the best step found so far until no improvement is obtained or a total of 20 trials have been done.

m is initially as follows:

(i) if $E(\mathbf{u}^o + \delta\mathbf{u}^{T_1}) \geqslant E(\mathbf{u}^o)$; then $m = 0.5$

(ii) if $E(\mathbf{u}^o + \delta\mathbf{u}^{T_1}) < E(\mathbf{u}^o)$ and the estimate $\delta\mathbf{u}$ did not violate the constraints; then $m = 1.414$

(iii) otherwise $m = 0.707$.

In case (ii) if the second trial is not an improvement, m is replaced by its reciprocal. Otherwise m is the same throughout the search. If a value of $m > 1$ leads to the constraints being violated then the step is modified as for the first trial. Violation of the constraints will stop the search process (unless it occurs in a second trial which is not an improvement for which case $1/m$ is then tried anyway).

Denoting the final best step by $\delta\mathbf{u}^F$, the base point for the next stage is $\mathbf{u} + \delta\mathbf{u}^F$.

It may happen that unless constrained in some way, one of the perturbations used for evaluating $\mathbf{A}$ may give rise to a point $\mathbf{u}$ where $\varepsilon(\mathbf{u})$ is indeterminate. The above constraints are not applied in this case and some of the end points of the confidence ellipsoid may indeed end up outside the constraint region. Instead, a maximum perturbation modulus "dumax" is specified and if equation (19) demands that some perturbations $\mathbf{du}$ have moduli $|\mathbf{du}|$ (which equal $(\lambda_j)^{-1/2}|\mathbf{dv}|$) which exceed this limit, then the value of F_α used is replaced by F'_α, say, which is such that the largest perturbation modulus equals "dumax". Note that the original value F_α is not overwritten. It is always used if possible and, in particular, the confidence limits described below refer to the original value F_α.

The iterative process will finish either when the maximum number of iterations specified have been performed, or the process has "converged". The process is deemed to have converged if for the last stage

(i) The values of the function E at all the perturbation points are greater than the base point, and

(ii) The final improvement in E for the stage $[E(\mathbf{u}^o) - E(\mathbf{u}^o + \delta\mathbf{u}^F)]$ is less than

$$(\text{conv})\,\hat{E}\left(\frac{N}{U-N}\right)F_\alpha(N, U-N)$$

where $\hat{E} = E(\mathbf{u}^o + \delta\mathbf{u}^F)$,

F is the value specified by the user, and

"conv" is a constant specified by the user.

The test (ii) arises from the fact that (on the linear theory) the difference between the value of $\hat{E}$ and the value of E evaluated on the boundary of the confidence region is

$$\hat{E}\left(\frac{N}{U-N}\right)F_\alpha(N, U-N)$$

It seems reasonable that some small fraction of this should be used as a convergence criterion. The value conv = 0.001 has been found successful in practice.

If the iterations cease because of convergence then information about the confidence region is printed. This output is not obtained if nonlinest stops because of the limit on the number of iterations.

To test the validity of the assumptions of linearity implied by equation (1), the values of the function E at the perturbation points may be compared with the value given by the linear theory, which is

$$\hat{E}\left(1 + \frac{N}{U-N}\right)F_\alpha(N, U-N)$$

If the final stage has used $F'_\alpha \neq F_\alpha$ for the perturbations then the value in the above expression is of course F'_α. In this case the value of F'_α is printed out as well as the above estimate and the actual values of E.

Note that in the final stage, the perturbations are to the ends of the ellipsoid of the *previous* stage. If the final correction is small, this will not matter. The last stage may be thought of as one giving the confidence region rather than a final correction to the estimate $\hat{\mathbf{u}}$.

The covariance matrix of the estimate $\hat{\mathbf{u}}$ is

$$\hat{\sigma}^2\mathbf{R}$$

where $\hat{\sigma}^2 = \dfrac{\hat{E}}{U-N}$, $\mathbf{R} = \mathbf{c}^{-1}$.

It follows that the correlations between the estimates u_ι and u_j are

$$r_{\iota j} = \frac{R_{\iota j}}{\sqrt{(R_{\iota\iota} R_{jj})}}$$

These correlations are printed out.
The confidence limits for the individual estimates (independently) are

$$\hat{u}_\iota \pm \delta u_\iota$$

where

$$\delta u_\iota = \sqrt{\left(\frac{N}{U-N}\right) \hat{E} F_\alpha R_{\iota\iota}}$$

A further useful indication of the nature of the confidence region may be obtained by considering the confidence limits for each estimate supposing that the other estimates are, in fact, exact, that is, a conditional confidence limit.
These limits are

$$\hat{u}_\iota \pm \delta u'_\iota$$

where

$$\delta u'_\iota = \sqrt{\left(\frac{N}{U-N}\right) \frac{\hat{E} F_\alpha}{c_{\iota\iota}}}$$

The geometrical interpretation is that the tangent-planes to the ellipsoid with normals in direction u_ι are at a distance δu_ι from the centre of the ellipsoid, and that the axis ι intercepts the ellipsoid at points $\delta u'_\iota$ from the centre.
Clearly

$$\delta u'_\iota \leqslant \delta u_\iota$$

The values of $\delta u'_\iota$ and δu_ι are printed out.
Since nonlinest requires an estimate of the eigensystem for each iteration, an initial matrix and vector must be provided by the user. However, this estimate may be very bad or even quite arbitrary and if so, ill-conditioning of c may occur; hence the facility has been provided of allowing an initial stage in which only the eigensystem of c is evaluated. (This does not, of course, involve the evaluation of c^{-1}.)

This facility is initiated by the boolean parameter "start". If "start" is set to *true*, the matrix is *automatically* set to the unit matrix and is still the matrix Q used in the transformation, but instead of the vector of eigenvalues we specify the vector whose elements are the actual perturbations to be employed along the axes of Q. (This vector is the vector diag. $[L^{-1}]$, thus defining L, and in this case $|\mathbf{dv}| = 1$.) The first stage then merely evaluates the eigensystem, subsequent stages being as fully described above. If "start" is set to *false*, then the matrix and vector are taken to be estimates of the eigensystem as if from a previous stage, *both* the matrix *and* the vector must be initially set by the user and a full iteration is then carried out.

3. Error detection facilities

Certain checks have been built into nonlinest. In most cases failure to satisfy the check results in exit from the procedure.
The checks are:

(i) If the constraint condition

$$ul_\iota \leqslant u^0_\iota \leqslant uu_\iota$$

is not satisfies on entry, the caption "initial point violates constraints" is printed. Exit.

(ii) The matrix $\mathbf{c} = \mathbf{a}'\mathbf{a}$ is positive definite and so has positive eigenvalues. The eigenvalues are checked to ensure that:
   (a) They are all positive.
   (b) The last one is the smallest. (The eigenvalues are calculated by means of the procedure SYMEIGEN which should produce them in descending order.)
   (c) The ratio of the smallest to the largest is not less than *b*, a parameter of the procedure.
   If any of these checks fail, the caption "eigenvalues not admissible" is printed. This is followed by the residual vector $\boldsymbol{\varepsilon}(\mathbf{u}^0 + \boldsymbol{\delta}\mathbf{u}^F)$ if called for. Exit.

(iii) If the search along the line joining the points $\mathbf{u}^0$ and $\mathbf{u}^0 + \boldsymbol{\delta}\mathbf{u}$ has not been successful after 20 trials, then nonlinest is considered to have failed. The

caption "fail-20 steps" is printed followed by the residual vector $(\mathbf{u}^o + \delta\mathbf{u}^B)$ if called for. Exit.

Reference

[1] H.H.Rosenbrock, Computer J. 3 (1960) 175.