PRACTICAL GUIDELINES Chapter 8[1]

- Quality of data (detection of outliers)
- Generation of initial guesses
- Overstepping
- Ill-conditioning
- Use of prior information

1. Quality of data (Inspection of data. Perform v*isual inspection* to spot potential *outliers*. **Outliers are data which appear to be *inconsistent* with the rest.** Furthermore, examine the residuals and if they are bigger than 3 or 4 standard deviations this is an indication of an outlier

A common sense approach for the analysis of engineering data:

(i)) Identify "first" set of potential outliers by visual observation of data
(iii) Eliminate potential outliers due to non-statistical reasons
(iv) Estimate the parameters and obtain the values of response variables and calculate the residuals.
(v) Plot the residuals and examine whether there any data point lie beyond 3 standard deviations around the mean response estimated by the model.
(vi) Examine whether the outliers should be discarded due to non-statistical reasons and eliminate them from the data set.
(vii) For all remaining "outliers" examine their effect on the model response and the estimated parameter values. This is done by performing the parameter estimation once with the outlier included in the data set and once without.
(viii) Determine which of the outliers are highly informative compared to the rest data points. The remaining outliers should have little or no effect on the model parameters or the mean estimated response of the model and hence, they can be ignored.
(ix) Prepare a final report where the effect of highly informative outliers is clearly presented.
(x) If possible, perform replicate experiments at the experimental conditions where the outliers were detected.

[1] Englezos, P. and N. Kalogerakis, "*Applied Parameter Estimation for Chemical Engineers*", Marcel-Dekker, New York, 2001 (ch. 8)

2. Generation of Initial Guesses

A good initial guess (starting values of the parameter vector) facilitates quick convergence of any iterative method and particularly of the Gauss-Newton method to the optimum parameter values. There are several approaches that can be followed

Nature and structure of the model: The nature of the mathematical model that describes a physical system may dictate a range of acceptable values for the unknown parameters. Furthermore, repeated computations of the response variables for various values of the parameters and subsequent plotting of the results provides valuable experience to the analyst about the behavior of the model and its dependency on the parameters. As a result of this exercise, we often come up with fairly good initial guesses for the parameters. The disadvantage of this approach is that it could be time consuming.

Asymptotic Behavior of the Model Equations: Quite often the asymptotic behavior of the model can aid us in determining sufficiently good initial guesses. For example,

Michaelis-Menten kinetics for enzyme catalyzed reactions,

$$y_i = \frac{k_1 x_i}{k_2 + x_i} \qquad (8.1)$$

When x_i tends to zero, the model equation reduces to

$$y_i \approx \frac{k_1}{k_2} x_i \qquad (8.2)$$

and hence, from the very first data points at small values of x_i, we can obtain k_1/k_2 as the slope of the straight line $y_i = \beta x_i$. Furthermore, as x_i tends to infinity, the model equation reduces to

$$y_i \approx k_1 \qquad (8.3)$$

and k_1 is obtained as the asymptotic value of y_i at large values of x_i.

Let us also consider the following exponential decay model, often encountered in analyzing environmental samples,

$$y_i = k_1 + k_2 \, exp(-k_3 x_i) \qquad (8.4)$$

At large values of x_i, (i.e., as $x_i \rightarrow \infty$) the model reduces to

$$y_i \approx k_1 \qquad (8.5)$$

whereas at values of x_i, near zero (i.e., as $x_i \rightarrow 0$) the model reduces to

$$y_i = k_1 + k_2(1 - k_3 x_i) \qquad (8.6a)$$

or

$$y_i = k_1 + k_2 - k_2 k_3 x_i \qquad (8.6b)$$

which is of the form $y_i = \beta_0 + \beta_1 x_i$ and hence, by performing a linear regression with the values of x_i near zero we obtain estimates of k_1+k_2 and k_2k_3. Combined with our estimate of k_1 we obtain starting values for all the unknown parameters.

Transformation of the Model Equations. The integrated form of substrate utilization in an enzyme catalyzed batch bioreactor is given by

$$\boldsymbol{k_1(x_i - x_0) = y_0 - y_i + k_2 ln\left(\frac{y_0}{y_i}\right)} \qquad (8.7)$$

where x_i is time and y_i is the substrate concentration. The initial conditions (x_0, y_0) are assumed to be known precisely. This is an implicit model (implicit in y_i).

Initial guesses for the two parameters can be obtained by noticing that this model is transformably linear since Equation 8.7 can be written as

$$\frac{y_0 - y_i}{x_i - x_0} = k_1 - k_2\left(\frac{1}{x_i - x_0}\right) ln\left(\frac{y_0}{y_i}\right) \qquad (8.8)$$

which is of the form $Y_i = \beta_0 + \beta_1 X_i$ where

$$Y_i = \frac{y_0 - y_i}{x_i - x_0} \qquad (8.9a)$$

and

$$X_i = \left(\frac{1}{x_i - x_0}\right) ln\left(\frac{y_0}{y_i}\right) \qquad (8.9b)$$

Initial estimates for the parameters can be readily obtained using linear least squares estimation with the transformed model.

The famous Michaelis-Menten kinetics expression shown below

$$y_i = \frac{k_1 x_i}{k_2 + x_i} \qquad (8.10)$$

can become linear by the following transformations also known as

Lineweaver-Burk plot: $$\left(\frac{1}{y_i}\right) = \frac{1}{k_1} + \frac{k_2}{k_1}\left(\frac{1}{x_i}\right) \qquad (8.11)$$

Eadie-Hofstee plot: $$y_i = k_1 - k_2\left(\frac{y_i}{x_i}\right) \qquad (8.12)$$

and *Hanes* plot: $$\left(\frac{x_i}{y_i}\right) = \frac{1}{k_1} x_i + \frac{k_2}{k_1} \qquad (8.13)$$

All the above transformations can readily produce initial parameter estimates for the kinetic parameters k_1 and k_2 by performing a simple linear regression.

Another class of models that are often transformably linear arise in heterogeneous catalysis. For example, the rate of dehydrogeneration of ethanol into acetaldehyde *over a Cu-Co catalyst*

$$y_i = \frac{k_5\left(x_{1i} - \frac{x_{2i}x_{3i}}{K_{eq}}\right)}{\left(1 + k_1x_{1i} + k_2x_{2i} + k_3x_{3i} + k_4x_{4i}\right)^2} \quad (8.14)$$

where y_i is the measured overall reaction rate and x_1, x_2, x_3 and x_4 are the partial pressures of the chemical species.

Good initial guesses for the unknown parameters, $k_1,\ldots,k_5$, can be obtained by linear least squares estimation of the transformed equation,

$$\sqrt{\frac{x_{1i} - \frac{x_{2i}x_{3i}}{K_{eq}}}{y_i}} = \frac{1}{\sqrt{k_5}} + \frac{k_1}{\sqrt{k_5}}x_{1i} + \frac{k_2}{\sqrt{k_5}}x_{2i} + \frac{k_3}{\sqrt{k_5}}x_{3i} + \frac{k_4}{\sqrt{k_5}}x_{4i} \quad (8.15)$$

which is of the form $Y = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4$.

Direct Search Approach

If we have very little information about the parameters, direct search methods, like the LJ optimization present an excellent way to generate very good initial estimates for the Gauss-Newton method.

Note that for algebraic equation models, direct search methods can be used to determine the optimum parameter estimates quite efficiently. However, if estimates of the uncertainty in the parameters are required, use of the Gauss-Newton method is strongly recommended, even if it is only for a couple of iterations.

OVERSTEPPING

Often in gradient methods the length of the increment of the parameters is large and as a result, the value of the objective function at the new parameter estimates could actually be higher than its value at the previous iteration.

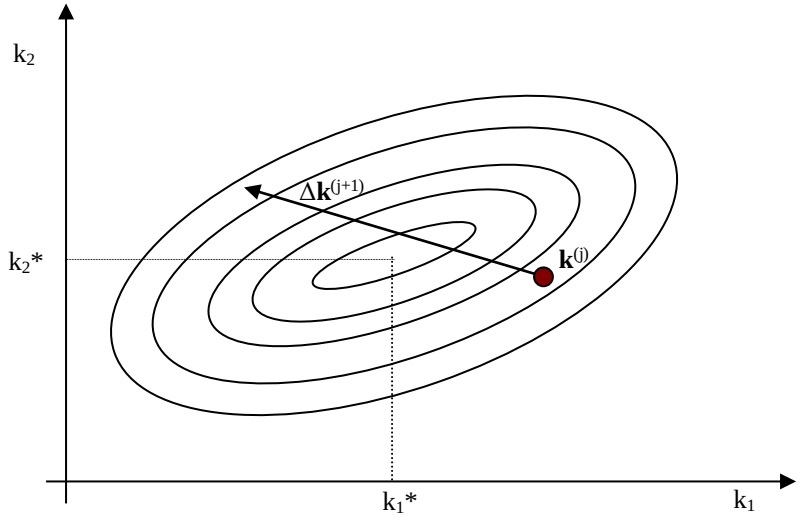

Figure 8.1 Contours of the objective function in the vicinity of the optimum. Potential problems with overstepping are shown for a two-parameter problem.

The classical solution to this problem is by limiting the step-length through the introduction of a stepping parameter, μ $(0<\mu\leq 1)$, namely

$$\mathbf{k}^{(j+1)} = \mathbf{k}^{(j)} + \mu\Delta\mathbf{k}^{(j+1)} \qquad (8.17)$$

The easiest way to arrive at an acceptable value of μ, μ_a, is by employing the *bisection rule* as previously described. Namely, we start with $\mu=1$ and we keep on halving μ until the objective function at the new parameter values becomes less than that obtained in the previous iteration, i.e., we reduce μ until

$$S(\mathbf{k}^{(j)}+\mu_a\Delta\mathbf{k}^{(j+1)}) < S(\mathbf{k}^{(j)}). \qquad (8.18)$$

Normally, we stop the step-size determination here and we proceed to perform another iteration of Gauss-Newton method.

ILL CONDITIONING **matrix A** (p x p matrix)

If two or more of the unknown parameters are highly correlated, or one of the parameters does not have a measurable effect on the response variables, matrix **A** may become singular or near-singular.

We have the so called *ill-posed* problem and matrix **A** is *ill-conditioned*.

A measure of the degree of ill-conditioning of a *nonsingular square* matrix is the *condition number*[2] (always greater than one). The *condition number* is equal to the square root of the ratio of the largest to the smallest *singular value* of **A**.

In parameter estimation applications, **A** is a positive definite *symmetric* matrix and hence, the *cond*(**A**) is also equal to the ratio of the largest to the smallest *eigenvalue* of **A**, i.e.,

$$cond(\mathbf{A}) = \frac{\lambda_{max}}{\lambda_{min}} \qquad (8.28)$$

In general

IF *condition number* $< 10^3$ ➔ the problem is well-posed.

IF *condition number* $> 10^{10}$ the problem is relatively ill-conditioned

IF *condition number* $> 10^{30}$ the problem is very ill-conditioned and we may encounter computer overflow problems.

NOTES:

(a) TRULY Ill-CONDITIONED PROBLEMS. If matrix **A** is ill-conditioned at the optimum (i.e., at $\mathbf{k}=\mathbf{k}^*$), there is not much we can do. This is a "truly" ill-conditioned problem and the estimated parameters will have questionable values with unacceptably large estimated variances.

In such a case it is best to re-examine the model. *Sequential experimental design* techniques can be helpful (see Chapter 12).

(b) ILL-CONDITIONING away from the OPTIMUM. If matrix **A** is reasonably well-conditioned at the optimum, **A** could easily be ill-conditioned when the parameters are away from their optimal values. This is often the case in parameter estimation.

In such cases, we may use: (i) a pseudoinverse[3] and/or (ii) Levenberg-Marquardt's modification.

---

[2] Golub G.H., and C.F. van Loan, Matrix Computations, John Hopkins University Presss, 1989 pp 79-81

[3] Englezos, P. and N. Kalogerakis, "*Applied Parameter Estimation for Chemical Engineers*", Marcel-Dekker, New York, 2001 (ch. 8)

Marquardt's Modification[4]

In order to improve the convergence characteristics and robustness of the Gauss-Newton method, Levenberg and later Marquardt proposed to add a small positive number, γ^2, to the diagonal elements of **A**.

Thus, the increment in the parameter vector is obtained by solving

$$(\mathbf{A} + \gamma^2\mathbf{I}\,)\,\Delta\mathbf{k}^{(j+1)} = \mathbf{b} \qquad (8.33)$$

If we consider the eigenvalue decomposition of **A**, $\mathbf{V}^T\mathbf{\Lambda}\mathbf{V}$ we have,

$$\mathbf{A}+\gamma^2\mathbf{I} = \mathbf{V}^T\mathbf{\Lambda}\mathbf{V}+\gamma^2\mathbf{V}^T\mathbf{V} = \mathbf{V}^T(\mathbf{\Lambda}+\gamma^2\mathbf{I})\mathbf{V} = \mathbf{V}^T\mathbf{\Lambda}^M\mathbf{V} \qquad (8.34)$$

where

$$\mathbf{\Lambda}^M = diag\,(\lambda_1+\gamma^2, \lambda_2+\gamma^2,\ldots,\lambda_p+\gamma^2) \qquad (8.35)$$

Thus all the *eigenvalues* of matrix **A** are all increased by γ^2, i.e., the eigenvalues are now $\lambda_1+\gamma^2, \lambda_2+\gamma^2,\ldots,\lambda_p+\gamma^2$.

Obviously, the large eigenvalues will be hardly changed whereas the small ones that are much smaller than γ^2 become essentially equal to γ^2 and hence, the condition number of matrix **A** is reduced from $\lambda_{max}/\lambda_{min}$ to

$$cond(\mathbf{A}) = \frac{\lambda_{max}+\gamma^2}{\lambda_{min}+\gamma^2} \approx \frac{\lambda_{max}}{\gamma^2} \qquad (8.36)$$

Scaling of Matrix A. When the parameters differ by more than one order of magnitude, matrix **A** may appear to be ill-conditioned even if the estimation problem is well-posed. The *reduced sensitivity coefficients* are introduced[5]

$$G_{Rij} = \left(\frac{\partial x_i}{\partial k_j}\right) k_j \qquad (8.37)$$

and the *reduced parameter sensitivity matrix*, $\mathbf{G}_R$, is related to our usual matrix, **G**, as follows

$$\mathbf{G}_R = \mathbf{G}\,\mathbf{K} \qquad (8.38)$$

Where. $\mathbf{K} = diag(k_1, k_2, \ldots, k_p)$ (8.39)

---

[4] Levenberg, K., "A Method for the Solution of Certain Non-linear Problems in Least Squares", *Quart. Appl. Math.*, II(2), 164-168 (1944). Marquardt, D.W., "An Algorithm for Least-Squares Estimation of Nonlinear Parameters", *J. Soc. Indust. Appl. Math.*, 11(2), 431-441 (1963).

[5] Englezos, P. and N. Kalogerakis, "*Applied Parameter Estimation for Chemical Engineers*", Marcel-Dekker, New York, 2001 (ch. 8)

USE OF "PRIOR" INFORMATION. Under certain conditions we may have some prior information about the parameter values. This information is often summarized by assuming that each parameter is distributed normally with a given mean and a small or large variance depending on how trustworthy our prior estimate is. The *Bayesian objective function*, $S_B(\mathbf{k})$, that should be minimized for algebraic equation models is

$$S_B(\mathbf{k}) = \sum_{i=1}^{N} [\hat{\mathbf{y}}_i - \mathbf{f}(\mathbf{x}_i,\mathbf{k})]^T \mathbf{Q}_i [\hat{\mathbf{y}}_i - \mathbf{f}(\mathbf{x}_i,\mathbf{k})] + (\mathbf{k}-\mathbf{k}_B)^T \mathbf{V}_B^{-1} (\mathbf{k}-\mathbf{k}_B) \quad (8.46)$$

and for differential equation models it takes the form,

$$S_B(\mathbf{k}) = \sum_{i=1}^{N} [\hat{\mathbf{y}}_i - \mathbf{y}(t_i,\mathbf{k})]^T \mathbf{Q}_i [\hat{\mathbf{y}}_i - \mathbf{y}(t_i,\mathbf{k})] + (\mathbf{k}-\mathbf{k}_B)^T \mathbf{V}_B^{-1} (\mathbf{k}-\mathbf{k}_B) \quad (8.47)$$

We have assumed that the prior information can be described by the multivariate normal distribution, i.e., $\mathbf{k}$ is normally distributed with mean $\mathbf{k}_B$ and covariance matrix $\mathbf{V}_B$.

The required modifications to the Gauss-Newton algorithm presented in Chapter 4 are rather minimal. At each iteration, we just need to add the following terms to matrix $\mathbf{A}$ and vector $\mathbf{b}$,

$$\mathbf{A} = \mathbf{A}_{GN} + \mathbf{V}_B^{-1} \quad (8.48)$$

and

$$\mathbf{b} = \mathbf{b}_{GN} - \mathbf{V}_B^{-1}(\mathbf{k}^{(j)} - \mathbf{k}_B) \quad (8.49)$$

where $\mathbf{A}_{GN}$ and $\mathbf{b}_{GN}$ are matrix $\mathbf{A}$ and vector $\mathbf{b}$ for the Gauss-Newton method as given in Chapter 4 for algebraic or differential equation models. The prior covariance matrix of the parameters ($\mathbf{V}_B$) is often a diagonal matrix and since in the solution of the problem only the inverse of $\mathbf{V}_B$ is used, it is preferable to use as input to the program the inverse itself.

From a computer implementation point of view, this provides some extra flexibility to handle simultaneously parameters for which we have some prior knowledge and others for which no information is available. For the latter we simply need to input *zero* as the inverse of their prior variance.

Practical experience has shown that (i) if we have a relatively large number of data points, the prior has an insignificant effect on the parameter estimates (ii) if the parameter estimation problem is ill-posed, use of "prior" information has a stabilizing effect. As seen from Equation 8.48, all the eigenvalues of matrix $\mathbf{A}$ are increased by the addition of positive terms in its diagonal. It acts almost like Marquadt's modification as far as convergence characteristics are concerned.

SELECTION OF WEIGHTING MATRIX Q IN LEAST SQUARES ESTIMATION

As we mentioned in Chapter 2, the user specified matrix $\mathbf{Q}_i$ should be equal to the inverse of $COV(\mathbf{e}_i)$. However, in many occasions we have very little information about the nature of the error in the measurements. In such cases, we have found it very useful to use $\mathbf{Q}_i$ as a normalization matrix to make the measured responses of the same order of magnitude. If the measurements do not change substantially from data point to data point, we can use a constant $\mathbf{Q}$. The simplest form of $\mathbf{Q}$ that we have found adequate is to use a diagonal matrix whose j^{th} element in the diagonal is the inverse of the squared mean response of the j^{th} variable,

$$Q_{jj} = \frac{1}{\left(\frac{1}{N}\sum_{i=1}^{N} \hat{y}_{j,i}\right)^2} \qquad (8.50)$$

This is equivalent to assuming a constant standard error in the measurement of the j^{th} response variable, and at the same time the standard errors of different response variables are proportional to the average value of the variables. This is a "safe" assumption when no other information is available, and least squares estimation pays equal attention to the errors from different response variables (e.g., concentration, versus pressure or temperature measurements).

If however the measurements of a response variable change over several orders of magnitude, it is better to use the non-constant diagonal weighting matrix $\mathbf{Q}_i$ given below

$$Q_{i,jj} = \frac{1}{\left(\hat{y}_{j,i}\right)^2} \qquad (8.51)$$

This is equivalent to assuming that the standard error in the i^{th} measurement of the j^{th} response variable is proportional to its value, again a rather "safe" assumption as it forces least squares to pay equal attention to all data points.

IMPLEMENTATION GUIDELINES FOR ODE MODELS

The issues that one needs to address more carefully in order to enhance the performance (robustness) of the Gauss-Newton method are (i) numerical instability during the integration of the state and sensitivity equations, (ii) ways to enlarge the region of convergence. In order to deal with numerical instability it is recommended to scale the state equations according to the order of magnitude of the given measurements, whenever possible. *Scaling* refers to the following: Instead of using variable xi we employ x/x_{max} or x/x_{avg}.

For example, when temperature ($T=x_1$) varies between 0 and 1000 K and concentrations $C=x_2$ vary between 0-5 mol/L then the two state variables will be of the same order of magnitude after scaling[6].

Increasing the Region of Convergence. A well-known problem of the Gauss-Newton method is its relatively small *region of convergence*. Unless the initial guess of the unknown parameters is in the vicinity of the optimum, divergence may occur. In order to increase the region of convergence we employ the *Information Index*[7] for each parameter, defined as

$$I_j(t) = k_j\left(\frac{\partial \mathbf{y}^T}{\partial k_j}\right)\mathbf{Q}\left(\frac{\partial \mathbf{y}}{\partial k_j}\right)k_j \quad ; \quad j=1,\ldots,p \qquad (8.67)$$

or equivalently, using the *sensitivity coefficient matrix*,

$$I_j(t) = k_j\boldsymbol{\delta}_j^T\mathbf{G}^T(t)\mathbf{C}^T\mathbf{Q}\,\mathbf{C}\,\mathbf{G}(t)\,\boldsymbol{\delta}_j k_j \quad ; \quad j=1,\ldots,p \qquad (8.68)$$

where $\boldsymbol{\delta}_j$ is a *p-dimensional* vector with 1 in the j^{th} element and zeros elsewhere.

The scalar $I_j(t)$ should be viewed as an index measuring the overall sensitivity of the output vector to parameter k_j at time t.

Thus given an initial guess for the parameters, we can integrate the state and sensitivity equations and compute the *Information Indices*, $I_j(t)$, j=1,…,p as functions of time.

Subsequently by plotting $I_j(t)$ versus time, we can spot where they become excited and large in magnitude. If observations are not available within this time interval, artificial data can be generated by *data smoothing* and *interpolation* to provide the missing sensitivity information. At the last iteration all artificial data are dropped and only the given data are used.

Example. *Pyrolytic dehydrogenation of benzene to diphenyl and triphenyl* (introduced first in Section 6.5.2). In Figure 8.2 the *Information Indices* are presented in graphical form with parameter values (k_1=355,400 and k_2=403,300). These values are three orders of magnitude away from the optimum. With this initial guess the Gauss-Newton method fails to converge as all the available sensitivity information falls outside the range of the given measurements.

---

[6] Englezos, P. and N. Kalogerakis, "*Applied Parameter Estimation for Chemical Engineers*", Marcel-Dekker, New York, 2001 (ch. 8)

[7] Kalogerakis, N., and R. Luus, "Improvement of Gauss-Newton Method for Parameter Estimation through the Use of Information Index", *Ind. Eng. Chem. Fundam.*, 22, 436-445 (1983b).

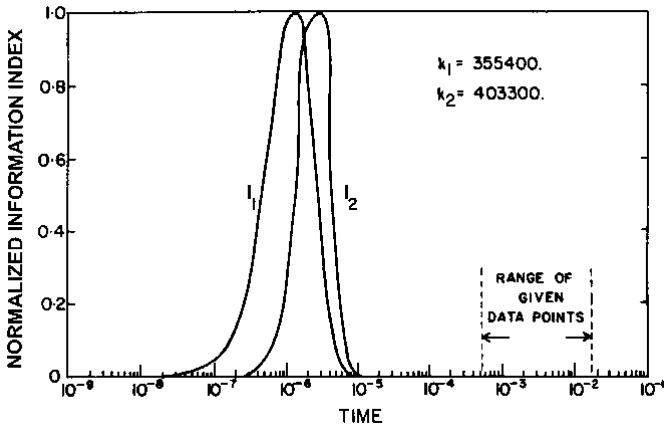

Figure 8.2 Normalized Information Index of k_1 and k_2 versus time to determine the best section of data to be used by the Gauss-Newton method (I_{1max}=0.0884, I_{2max}=0.0123) [reprinted from Industrial Engineering Chemistry Fundamentals with permission from the American Chemical Society].

We mow generate artificial data by interpolation and then use of the Gauss-Newton method brings the parameters to the optimum (k_1=355.4 and k_2=403.3) in nine iterations.

As seen in Figure 8.3, in terms of experimental design the given measurements have been taken at proper times, although some extra information could have been gained by having a few extra data points in the interval [10^{-4}, 5.83×10^{-4}].

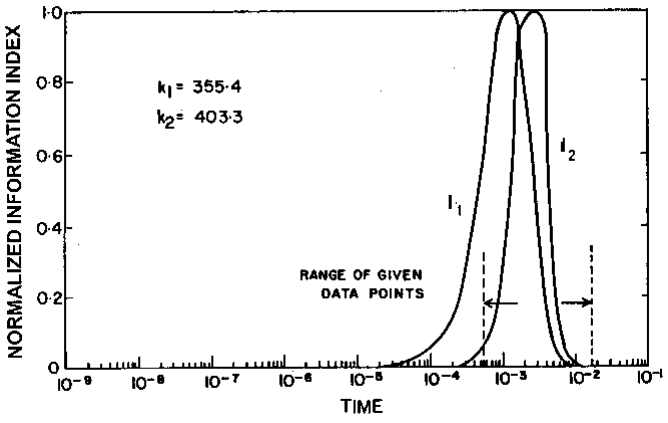

Figure 8.3 Normalized Information Index of k_1 and k_2 versus time to determine whether the measurements have been collected at proper times (I_{1max}=0.0885, I_{2max}=0.0123) [reprinted from Industrial Engineering Chemistry Fundamentals with permission from the American Chemical Society].

Use of Direct Search Methods

A simple procedure to overcome the problem of the small region of convergence is to use a two-step procedure whereby direct search optimization is used to initially to bring the parameters in the vicinity of the optimum, followed by the Gauss-Newton method to obtain the best parameter values and estimates of the uncertainty in the parameters (see chapter 8 section 8.7.2.3) .

For the homogeneous gas phase reaction of NO with O_2 (first presented in Section 6.5.1): In Figure 8.4 we see that the use of direct search (LJ optimization) can increase the overall size of the region of convergence by at least two orders of magnitude.

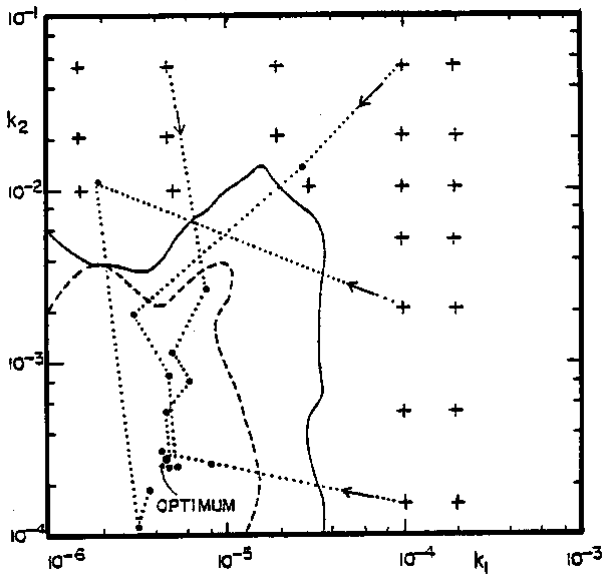

Figure 8.4Use of the LJ optimization procedure to bring the first parameter estimates inside the region of convergence of the Gauss-Newton method (denoted by the solid line). All test points are denoted by +. Actual path of some typical runs is shown by the dotted line.