JOURNAL OF COMPUTATIONAL PHYSICS 6, 367–377 (1970)

Diagnostic Least Squares*

R. F. CURL, JR.

Chemistry Department, Rice University, Houston, Texas 77001

Received December 29, 1969

In the determination of a set of parameters from a set of observations by the method of least squares, it often occurs that the relationships provided by the observations are not really linearly independent when the random errors in the observations are considered. A method for finding physically reasonable parameters and confidence limits for parameters is described. This method is based on parameter scaling and diagonalization of the matrix of the normal equations.

INTRODUCTION

Several authors [1–5] have used a technique for least squares analysis of experimental data which involves diagonalizing the matrix of the normal equations, thereby decoupling the equations which determine the parameters. This approach is useful when one or more of the parameters is poorly determined because of linear dependence. This paper is concerned with an extension of this approach which it is believed will be most useful.

Review of Least Squares

In least square analysis the usual situation is that a model is proposed in which a set of observations, y, (n in number) are thought to be known functions of a set of parameters, x, (m in number).

$$y_i = f_i(x_1, \ldots, x_m), \qquad i = 1, \ldots, n. \tag{1}$$

Because of experimental errors or inadequacies in the model (1) is not expected to hold exactly. There will be residuals, r_i, defined by

$$r_i = y_i - f_i(x_1, \ldots, x_m), \qquad i = 1, \ldots, n. \tag{2}$$

* This work was supported by Grant C-071 of the Robert A. Welch Foundation.

Least squaring then consists of finding the set of parameters, x, such that the sum of squares or residuals, V, is minimized,

$$V = \sum_{i=1}^{n} r_i^2. \tag{3}$$

In order to determine the parameters from the observations, the functional relationships between observations and parameters are linearized (if they are not already linear) by assuming a set of initial values, $x_i^{(0)}$, for the parameters and expanding in a Taylor series about the initial values keeping only the first two terms. This linearization leads to the expression

$$r_i = y_i - \left[f_i(x_1^{(0)},..., x_m^{(0)}) + \sum_j \frac{\partial f_i}{\partial x_j}\Big|_{x^{(0)}} (x_j - x_j^{(0)})\right]. \tag{4}$$

Defining the vectors r(length n), a(length n), d(length m) by

$$a_i = y_i - f_i(x_1^{(0)},..., x_m^{(0)}), \tag{5}$$

$$d_j = x_j - x_j^{(0)}, \tag{6}$$

and the matrix A $(n \times m)$

$$A_{ij} = \frac{\partial f_i}{\partial x_j}\Big|_{x^{(0)}}, \tag{7}$$

(4) becomes in matrix notation

$$r = a - Ad, \tag{8}$$

and the equation for V becomes

$$\begin{aligned} V &= \tilde{r}r \\ &= (\tilde{a} - \tilde{d}\tilde{A})(a - Ad). \end{aligned} \tag{9}$$

The conditions for minimization of V are

$$\frac{\partial V}{\partial d_j} = 0, \qquad j = 1,..., m. \tag{10}$$

This leads to the normal equations

$$Bd = b, \tag{11}$$

where $B = \tilde{A}A$ and $b = \tilde{A}a$. If the observations determine the parameters, B is nonsingular, and the changes in parameters are determined,

$$d = B^{-1}b. \tag{12}$$

An iterative solution for the parameters is employed, i.e., a new initial set, $x^{(1)}$, is found from

$$x^{(1)} = x^{(0)} + d, \tag{13}$$

and the Taylor series expansion is made about the new point $x^{(1)}$ and the process is repeated. If we are fortunate the iteration will converge and a least square set of parameters will be obtained.

The procedure described is unweighted least squares. Implicit in it is the assumption that the experimental error in each observation is the same. This might be described statistically by assuming that the actual observations are samples from a set of independent normal distributions with the same standard deviation σ. If the different observations y_i are expected to be samples from independent normal distributions with different standard deviations, σ_i, it is possible to take this into account by weighting each of the equations for the residuals r_i [Eq. (8)] by $1/\sigma_i$.

Linear Dependence and Diagonalization of the Normal Matrix

Suppose, however, that the matrix B is singular or nearly so. Then the step from (11) to (12) is not possible. The obvious conclusion is that the observations do not determine the parameters. However, the observations do determine something, i.e., there is some information contained in the observations which we would like to extract.

The matrix B is a real, symmetric matrix and can be diagonalized by an orthogonal transformation,

$$\lambda = TB\tilde{T}, \tag{14}$$

where λ is a diagonal matrix. If B is nonsingular, all the eigenvalues λ will be greater than zero. If B is singular, one or more of the eigenvalues will be zero. Defining $c = Tb$ and $D = Td$, (11) may be transformed to

$$\lambda D = c \tag{15}$$

and the equations are decoupled,

$$\lambda_1 D_1 = c_1, \text{ etc.} \tag{16}$$

Then all the D's for which the corresponding λ's are greater than zero can be found. These linear combinations of the original parameters are determined by the data.

There is, however, the problem that the size of eigenvalues depends on the scaling of the parameters. When the calculation and diagonalization of B is carried out numerically no eigenvalues will be exactly zero because of rounding errors. Lees [5] has considered the question of parameter scaling from the point of view of numerical accuracy and has developed a method for parameter scaling which makes the eigenvalue size truly reflect the linear dependence of the corresponding set of parameters.

Diagnostic Least Squares

Reasonable Sets of Parameters

The original parameters x have some fairly well defined physical significance, since they are part of the model which is intended to describe the experimental observations. Often the x parameters may be at least roughly estimated from physical considerations or by comparison to similar parameters determined for other cases.

On the other hand, the physical significance of the parameters D, which are determined from (16), is usually very obscure. Thus one is left in the rather unsatisfactory position of being unable to discuss the physical significance of the parameters (i.e., the D's) which can be obtained from the experiment.

It is always possible (by resorting, if necessary, to an opinion poll of experts in the field) to estimate a range for each of the x parameters such that if the parameter x_i falls outside the range $x_i^L < x_i < x_i^U$, one is surprised. If the parameter falls into the range its value is considered "reasonable." This range may be extremely large if little is known about x_i either theoretically or by analogy to similar situations. On the other hand this range may be quite narrow if much is known about similar situations.

The results of the procedure to be described will depend on the estimates of the ranges for the parameters and these estimates should always be explicitly stated so that the reader will be able to decide for himself whether they are reasonable.

Now in order to proceed it is necessary to cast what is meant by "reasonable" into statistical language. Unfortunately this step is quite arbitrary. The procedure adopted here is to call the interval x_i^L to x_i^U a normal distribution 90 % confidence interval. Thus the a priori best estimate of x_i, $x_i^{(0)}$ is given by

$$x_i^{(0)} = (x_i^L + x_i^U)/2, \tag{17}$$

with estimated standard deviation

$$\sigma_i = (x_i^U - x_i^L)/3.290. \tag{18}$$

A new set of parameters, z_i, can be defined,

$$z_i = (x_i - x_i^{(0)})/\sigma_i , \tag{19}$$

with the consequence that the z's are expected to be normally distributed with unit standard deviation and zero mean.

Now the z's are used instead of the d's in the least squares procedure. In order to distinguish this treatment quantities referring to the z parameters are primed. Thus $B'z = b'$, $\lambda' D' = c'$, $D' = T'z$, etc. The reader should keep in mind that tilde's instead of primes are being used for the matrix transpose.

The new parameters D' which result from the diagonalization process are determined by the equation

$$\begin{aligned} D_i' &= c_i'/\lambda_i', \qquad \lambda_i' \geqslant \sigma^2, \\ &= 0, \qquad \lambda_i' < \sigma^2 \end{aligned} \tag{20}$$

(in weighted least squares $\sigma = 1$ and $\lambda_i' \geqslant 1$ is the condition). It will be shown below that the linear combinations which are better determined by the observations than by the a priori estimates will have the corresponding $\lambda_i' \geqslant \sigma^2$. On the other hand for $\lambda_i' < \sigma^2$ the a priori estimates are expected to be more reliable. Now there is a value of D_i for each of the m values of i,

$$z = \tilde{T}' D', \tag{21}$$

and

$$x_i = x_i^{(0)} + z_i \sigma_i . \tag{22}$$

This procedure finds a set of parameters x_i which are fitted to the experimental data when the experimental data provides a more precise estimate than the guessed set (vide infra). Thus if

$$V < 2n\sigma^2, \tag{23}$$

where V is calculated using Eqs. (2) and (3) and the x's of (22), it is possible to say that the data can be fitted within experimental error by the model (Eq. (1)). The x's of (22) can be examined to see if they are physically reasonable. If they are, then it is possible to say that at least one set of parameters which fit the data and are physically reasonable exist. No claim for uniqueness is made.

Ordinary least squares as described in (12) is appropriate to the case in which $\lambda_i \geqslant \sigma^2$ for every value of i.

Confidence Intervals for the Parameters

The diagnostic least squares procedure would be of greater value if an estimate of confidence intervals for the parameters after the diagnostic least squares can be made.

Consider first how confidence intervals are estimated for parameters in ordinary least squares. The matrix of covariances, C, for the parameters can be found by a straight forward statistical treatment to be

$$C = \sigma^2 B^{-1}, \tag{24}$$

where $C_{ij} = \langle \Delta x_i \, \Delta x_j \rangle$. A normal distribution 95% confidence interval for parameter x_i after ordinary least squaring would be given by

$$x_i = (x_i + d_i) \pm 1.96 \sqrt{C_{ii}}. \tag{25}$$

Now in diagnostic least squares the transformation T which diagonalizes B does not alter the fundamental situation. The new parameters D are on a footing equivalent to the original parameters. Thus the matrix of covariances, C_D, for the D parameters is given by

$$C_D = \sigma^2 \lambda^{-1}. \tag{26}$$

Since λ is diagonal the matrix of covariances, C_D, is diagonal and the distributions of the parameters D are independent. It is clear that parameters with large λ are well determined while those with small λ are poorly determined.

The rationale for the procedure for finding reasonable sets of parameters becomes clearer. The range x_i^L to x_i^U for a parameter x_i is a guessed 90% confidence interval. These guessed values of x_i are assumed independent so that the guessed matrix of covariances for x is diagonal with elements

$$(C_x)_{ij} = \sigma_i^2 \delta_{ij}, \tag{27}$$

where $\sigma_i = 1/3.290[x_i^U - x_i^L]$. Occasionally it may be possible to correlate the errors in two more different x's. That is, it may be possible to guess a nondiagonal C_x. The procedure to be followed in this case is outlined in the Appendix.

Continuing with the matrix of covariances given by (27), the introduction of the parameters z gives a corresponding matrix of covariances C_z which is diagonal with all diagonal elements equal to one. When transformed to the D' basis, it remains the same, being the unit matrix. Then the procedure of using the experimental data to determine those D_i' for which the corresponding $\lambda_i' \geqslant \sigma^2$ and using the estimate to determine those D_i''s for which $\lambda_i' < \sigma^2$ (i.e., setting that $D_i' = 0$), means that one is taking the method of determining D_i' which gives the smaller covariance. The result is a hybrid matrix of covariances. For fitted D_i''s the diagonal elements are σ^2/λ_i'. For unfitted D_i''s the diagonal elements are one. All off-diagonal elements are zero. Call the hybrid matrix of covariances $C_D^{(h)}$.

The transformation which diagonalized B can be reversed, giving a new matrix of covariances for the z's, $C_z^{(h)}$,

$$C_z^{(h)} = \hat{T}' C_D^{(h)} T'. \tag{28}$$

The normal distribution 90% confidence interval for z_i can be stated

$$z_i = z_i \pm 1.645[(C_z^{(h)})_{ii}]^{1/2}, \tag{29}$$

and the 90% confidence interval for the x's can be given

$$x_i = x_i^{(0)} + z_i\sigma_i \pm 1.645\sigma_i[(C_z{}^h)_{ii}]^{1/2}. \tag{30}$$

This is to be contrasted with the same interval before least squaring,

$$x_i = x_i{}^0 \pm 1.645\sigma_i . \tag{31}$$

Comparison to Conditioning the Normal Equations by Including the Estimate

Another approach which has been used when the normal equations are linearly dependent is to include the estimates of the parameters, weighted by an appropriate factor, in the equations to be fitted. Thus to the n equations $Ad = a$, the m equations $w_i d_i = 0$ $(i = 1,..., m)$ are added. Presumably w_i should be inversely proportional to the estimated standard deviation of x_i,

$$w_i = w/\sigma_i . \tag{32}$$

In terms of the z parameters the equations to be used in least squaring are $A'z = a$ and $wz = 0$. Then the conditioned normal matrix, B_c', is given by

$$B_c' = B' + w^2 I, \tag{33}$$

where B' is the original normal matrix and I is the $m \times m$ unit matrix.

The same orthogonal transformation, T', which diagonalizes B' also diagonalizes B_c'. Thus

$$\lambda_c' = \lambda' + w^2 I \tag{34}$$

and

$$(D_c')_i = c_i'/(\lambda_i' + w^2). \tag{35}$$

If w is chosen to be very large $(w \gg \sigma)$, the parameters are not responsive to the data. If w is chosen to be small $(w \ll \sigma)$, the parameters with small dependence

on the data ($\lambda_i' \ll \sigma^2$) are fitted to the data even though they are not determined by the data. The optimum choice of w is probably $w \sim \sigma$.

When the choice $w = \sigma$ is made, the parameters, $(D_c')_i$, obtained will be increasingly unrelated to the data as λ_i' decreases.

The conditioning equation method corresponds to taking a sort of weighted average of the a priori and least squares decoupled parameters with the weighting factor varying with parameters. The diagnostic least squares method corresponds to deciding, on the basis of which gives the lower standard deviation, to take either the least squares or the a priori decoupled parameter. It is clear that for $\lambda_i' \sim \sigma$ there is a severe risk of making the wrong decision. In that case one should try both decisions and even report the result of both. Evidence obtained at some future time may then provide a basis for deciding which result is correct.

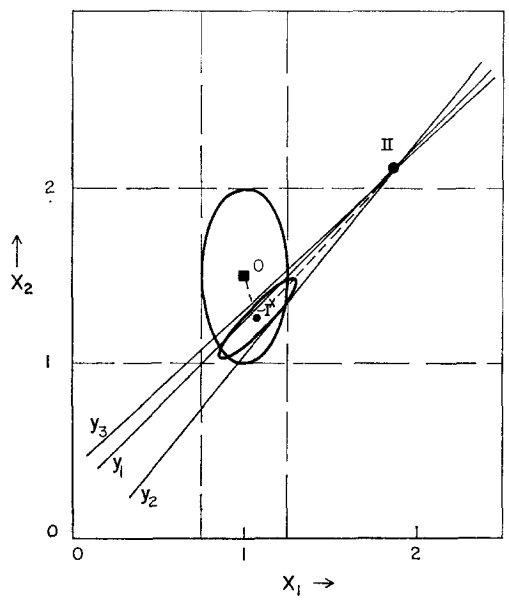

FIG. 1. A simple example of the method of diagnostic least squares. The point marked with a zero is the original best estimate of the parameters x_1 and x_2. The large ellipse about this point gives the original estimated standard deviations. The observations provide a relationship between the two parameters indicated by the lines marked y_1, y_2, and y_3. The point marked I is the new estimate of the parameters provided by the diagnostic least squares when the error in the observations is large ($\sigma = 5$). The ellipse around it gives the estimated standard deviations. The point marked II is the new estimate of the parameters provided by either diagnostic least squares or ordinary least squares when the error in the observations is small ($\sigma = 0.1$). The dashed line connecting the original best estimate to the point II is the locus of results obtained by the conditioning equation approach. The X on the line corresponds to the weight factor $w = 5$ ($w = \sigma$ when $\sigma = 5$).

A Simple Example

A simple example of the diagnostic least squares procedure is given in Table I and Fig. 1. In this example there are two parameters, x_1 and x_2, which are to be determined from three observations, y_1, y_2, and y_3. The large ellipse in Fig. 1 indicates the a priori estimated reasonable range of parameters. The result of ordinary least squares is indicated by the point marked II. This point is far

TABLE I

A Simple Example
Postulated functional relationship
$y_1 = -20.0x_1 + 20.0x_2$
$y_2 = 82.9x_1 - 68.8x_2$
$y_3 = -63.7x_1 + 69.3x_2$
Observations
$y_1 = 5$
$y_2 = 10$
$y_3 = 27$
Original estimates of parameters
$x_1 = 1.0 \pm 0.25$
$x_2 = 1.5 \pm 0.5$
Diagnostic Least squares
$\lambda_1 = 1{,}176.$ $\quad D_1 = -0.46\Delta z_1 + 0.88\Delta z_2$
$\lambda_2 = 3.5$ $\quad D_2 = 0.88\Delta z_1 + 0.46\Delta z_2$
Case I, $\sigma = 5$
$x_1 = 1.07 \pm 0.22$
$x_2 = 1.25 \pm 0.24$
$V = 129.$
Case II, $\sigma = 0.1$
$x_1 = 1.88 \pm 0.01$
$x_2 = 2.11 \pm 0.01$
$V = 0.06$

outside the reasonable range of x. If the data is quite accurate, as indicated by the low σ for Case II in Table I, this may be the correct result. The "unreasonableness" of x must be explained on physical grounds. However, if the data is not very accurate, as indicated for case I in Table I, then the "unreasonable" result for x_1 is almost certainly because of random error in the experimental data.

The diagnostic least-squares procedure indicates this by the value of λ_2 which is much smaller than σ^2. Further the diagnostic least squares procedure gives the parameters indicated by the point marked I on Fig. 1, with the 90% confidence limits indicated by the ellipse surrounding it.

With only two parameters it is possible to graphically illustrate the way in which the parameters depend on the observations. The information obtainable from diagnostic least squares is essentially the same as that obtainable from a cursory examination of the graph. The real power of the method becomes apparent when three or more parameters are involved and it is no longer possible to construct such graphs.

The result of the conditioning equations method is indicated by a line corresponding to the locus of all values of w. The point $w = 5$ (appropriate to $\sigma = 5$, Case I) is marked with an X.

Applications

It is possible to envisage many applications in real physical situations. The method offers the following information.

1. Which linear combinations of parameters are determined by the data and which are not. It is not necessary to have the data.

2. Whether a set of parameters which fit the data and are physically reasonable exist. The application in Ref. [4] is of this type.

3. Estimated standard deviations of the least squares parameters and correlations between parameters obtainable from the hybrid matrix of covariances.

4. A choice of several least squares sets of parameters one of which is likely to be nearly correct.

The references already cited [1, 2, 3, 4, 5] illustrate the application of at least part of the method. We are using the complete method for the analysis of the internal rotation structure of the microwave spectrum of methyl isocyanate.

Appendix

If a matrix of covariances C_x for the original parameters x may be estimated a priori, the diagnostic least squares proceeds as follows.

First C_x is diagonalized by a transformation R giving eigenvalues $\Lambda_i (\Lambda = RC_x\tilde{R})$. Then the new parameters z are defined by

$$z_i = \left(\sum_j R_{ij}x_j\right)\Big/(\Lambda_i)^{1/2}, \qquad i = 1,..., m. \tag{A1}$$

The matrix of covariances for the z's, C_z, is now just the unit matrix. The normal equations are set up as though the z's are the original parameters. B is diagonalized ($\lambda = TB\tilde{T}$, $D = Tz$). The estimated matrix of covariances for the parameters D is just a unit matrix. For all $\lambda_i \geqslant \sigma^2$, the experimental data determines the corresponding D_i more precisely than the estimate, while for $\lambda_i < \sigma^2$ the estimate determines D_i more precisely. Thus

$$\begin{aligned} D_i &= c_i/\lambda_i , \qquad \lambda_i \geqslant \sigma^2, \\ D_i &= 0, \qquad\quad\ \lambda_i < \sigma^2. \end{aligned} \tag{A2}$$

The hybrid matrix of covariances, C_D', is given by

$$\begin{aligned} (C_D')_{ij} &= (\sigma^2/\lambda_i)\,\delta_{ij} , \qquad \lambda_i \geqslant \sigma^2, \\ (C_D')_{ij} &= \delta_{ij} , \qquad\qquad\ \lambda_i < \sigma^2. \end{aligned} \tag{A3}$$

A set of improved z's and a hybrid C_z' may be found by transforming back,

$$\begin{aligned} z &= \tilde{T}D, \\ C_z' &= \tilde{T}C_D'T. \end{aligned} \tag{A4}$$

A new set of x's and a hybrid C_x' may be found by continuing the reverse transformation,

$$\begin{aligned} x_i &= x_i^{(0)} + \sum_j R_{ji}(\Lambda_j)^{1/2}\, z_j , \\ C_x' &= \tilde{R}\Lambda^{1/2}C_z'\Lambda^{1/2}R, \end{aligned} \tag{A5}$$

and confidence intervals for x may be found from the matrix of covariances C_x'.

References

1. R. Diamond, *Acta Crystallogr.* **11** (1958), 129; **19** (1965), 774.
2. P. D. Foster and R. F. Curl, *J. Chem. Phys.* **45** (1966), 3760.
3. M. J. Bruton and L. A. Woodward, *Spectrochim. Acta Part A*, **23** (1967), 175.
4. J. J. Keirns and R. F. Curl, *J. Chem. Phys.* **48** (1968), 3773.
5. R. M. Lees, *J. Mol. Spectrosc.* **33** (1970), 124.