The Efficient Use of Function Minimization in Non-linear Maximum-likelihood Estimation

By G. J. S. Ross

Statistics Department, Rothamsted Experimental Station, Harpenden

SUMMARY

Maximum-likelihood estimation problems can be solved numerically using function minimization algorithms, but the amount of computing required and the accuracy of the results depend on the way the algorithms are used. Attention to the analytical properties of the model, to the relationship between the model and the data, and to descriptive properties of the data can greatly simplify the problem, sometimes providing a method of solution on a desk calculator. This paper describes how parameter transformation, sequential minimization and nested minimization can be used to solve particular problems. Applications to well-known problems of distribution fitting, quantal responses and least-squares curve fitting are described. The implications for computer programming are discussed.

1. INTRODUCTION

IN this paper it will be argued that while general function minimization algorithms are extremely important for solving non-linear maximum-likelihood and least-squares estimation problems their value can be dramatically increased if they are used in a manner appropriate to a given problem. The context of this paper is a statistical advisory service in which many thousands of routine data-processing jobs are handled each year. There is no time to study each individual set of data and therefore all initial estimates of parameters must be computed by special routines appropriate to each model. Only when the program fails to converge is the data set referred to a statistician for comment.

This paper aims to show the common features of the special techniques which have been used successfully in a wide variety of models. The techniques, introduced in Section 1.3, are not all new but their systematic use in function minimization seems to have been neglected. Statisticians have often used *ad hoc* methods to obtain approximate estimates of parameters, and function minimization can be used to refine these estimates. The appropriate techniques for a given problem are found by a combination of *a priori* statistical reasoning and practical experience.

The methods discussed are valuable not only for models which occur commonly in routine work but also for "one-off" models such as the seven-parameter model discussed in Section 4.2 which failed to converge using standard minimization programs. An important requirement is that minimization programs should not be too rigid in their specification, and that the minimization sub-routines can be used with more than one function in any one use of the program. Programming considerations are discussed in Section 5.

For standard problems the ideal situation can be summarized as follows:

(i) the user need only supply the data and select the model;
(ii) convergence is as rapid and as accurate as possible;
(iii) insoluble cases are detected as early as possible;
(iv) if multiple solutions exist and are sufficiently distinct the intended solution will be found;
(v) appropriate statistical analysis is provided.

These conditions are not generally met by standard minimization programs and can only be achieved by special programming for individual models in the light of experience of a variety of data sets.

1.1. *Use of Minimization Algorithms*

Maximum-likelihood estimation problems may be solved formally by evaluating the negative log likelihood (subsequently called the likelihood) for trial values of the parameters and using efficient function minimization algorithms. This leads to the attractive proposition that a single computer program will be able to handle problems of great complexity with the minimum of special programming for each problem.

The most popular algorithms for unconstrained minimization of functions of several variables are the conjugate gradient method of Powell (1964), the orthogonal search method of Rosenbrock (1960) and the simplex method of Nelder and Mead (1965). If function derivatives are computable the conjugate gradient method of Fletcher and Powell (1963) may be used, and if the function is a sum of squares the methods of Powell (1965) and Marquardt (1963) are suitable. Good surveys of these methods are provided by Kowalik and Osborne (1968) and Box *et al.* (1969). These algorithms perform well for various problems yet there are many reports of failure to converge satisfactorily when fitting models to real data (e.g. Chapter 6 of Kowalik and Osborne, 1968).

An alternative approach is to take more care in presenting the likelihood function so that the minimization problem is as simple as possible. The programming problems are not necessarily much more elaborate than the simple evaluation of the likelihood, but there is an essential change of role. A minimization algorithm is a tool to assist in a statistical analysis, rather than a general system operating on a particular function. To operate on the likelihood alone is to ignore much useful information that can greatly improve the computing efficiency.

In this paper several methods are described that have enabled various problems to be solved rapidly and accurately using a quadratically convergent modification of Newton's method (Box *et al.*, 1969). Some of these problems could not be solved at all by direct minimization. To understand these methods we must first study the relationship between data, a statistical model and its likelihood function in parameter space.

1.2. *Data, Statistical Models and Likelihood Functions*

The likelihood function measures the goodness of fit of a statistical model to data for given values of the unknown parameters. Contours in parameter space represent members of the family of models that fit equally well. In particular a minimum (local or global) is surrounded by a contour that includes all members for which the likelihood is less than a certain value. Now unless the parameters happen to represent well-determined features of the data, the contours must be expected to be elongated, oblique to the axes and curved, giving the notorious "narrow descending valleys" that minimization algorithms are designed to negotiate.

As an example consider the data in Table 1.

TABLE 1

x	1	2	3	4	5
y	2	3	4	4	5

and suppose that the model

$$E(y) = b(1-r^x), \quad 0<r<1, \tag{1}$$

is to be fitted by least squares.

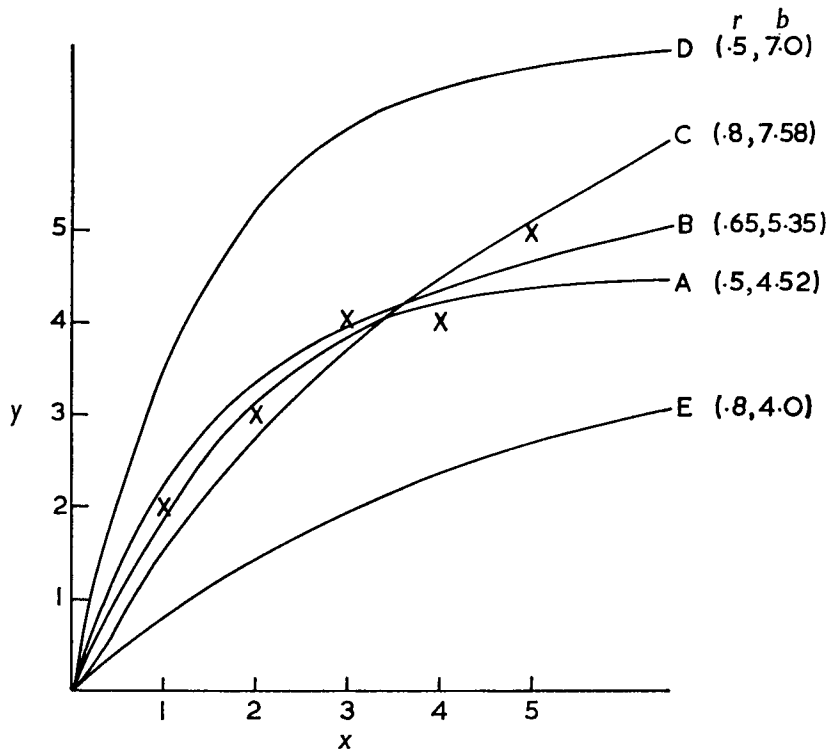


FIG. 1. Data of Table 1 with five curves of the family $y = b(1-r^x)$.

Specimen curves of the family are shown in Fig. 1, which show that curves A, B and C all fit quite well whereas curves D and E do not. Hence it is not surprising that the least-squares function, $R(r, b)$ displayed in Fig. 2, has a low contour $R = 1$ which includes the points corresponding to the parameter values of A, B and C, whereas the points D and E lie well outside this contour. The parameter b, the asymptote, is an extrapolated quantity that depends strongly on r, so that the rough outline of Fig. 2 can be predicted from study of Fig. 1 without any calculation. For example,

we may predict that adding a further data point (10, 5) will determine b accurately at around $b = 5$, which will mean a much flatter contour with very little interrelation between b and r.

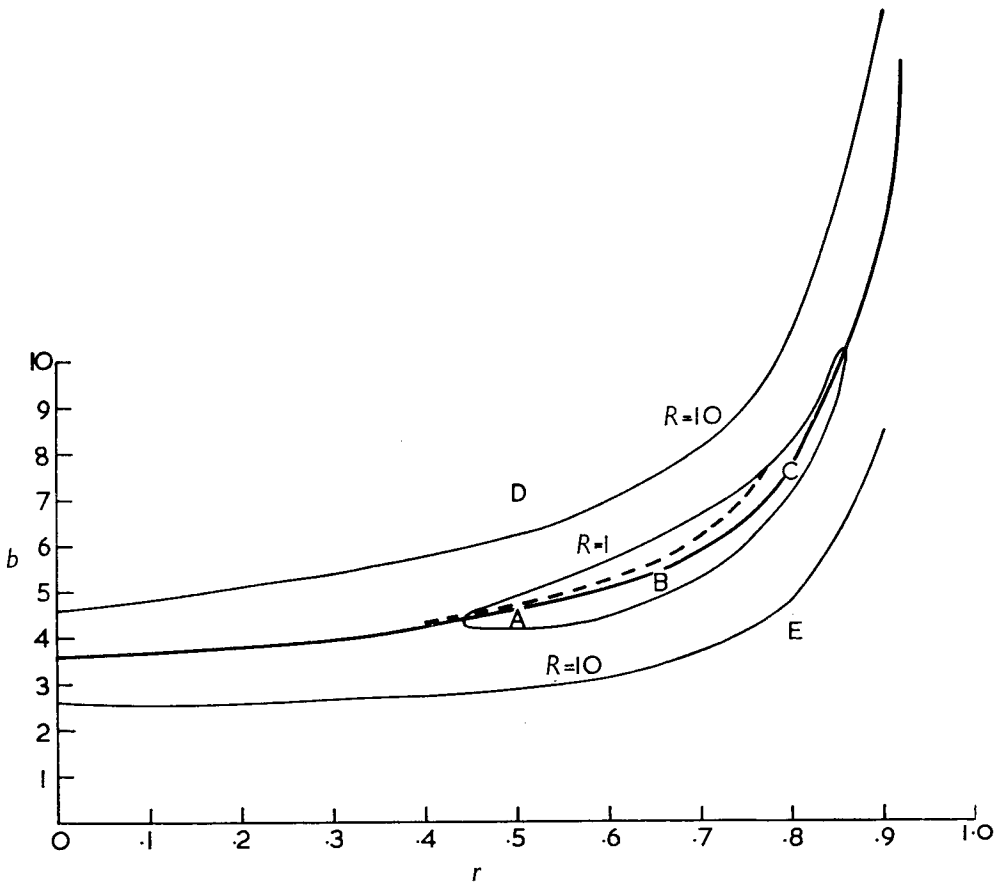


FIG. 2. Contours of the function $R(r, b)$, the residual sum of squares after fitting $E(y) = b(1-r^x)$ to the data of Table 1. The curve ABC gives for each r the value of b for which R is a minimum. The broken curve represents curves in Fig. 1 which pass through (3, 4).

This example shows that the data should not be ignored in specifying the likelihood. A minimization search of the function displayed in Fig. 2 is an unnecessarily elaborate way of detecting the existence of the long narrow valley, and an inefficient way of solving the estimation problem.

1.3. *Simplifying the Estimation Problem*

Minimization algorithms converge rapidly if the following objectives are attained:

(a) initial estimates are good;

(b) the likelihood function is well approximated by a quadratic in the neighbourhood of the minimum, in particular if the contour including the initial estimate is approximately an ellipsoid;

(c) the information matrix is well conditioned, which means that the parameter estimates are not strongly intercorrelated;
(d) the dimensions are few.

It has been customary to make the user responsible for the initial estimates, to regard the shape of the function as an unfortunate property of non-linear models, and hope that objectives (c) and (d) are sufficiently well met to allow the use of algorithms employing line minimization and orthogonal or conjugate gradients. However, some of these objectives can be better realized by re-formulating the problem.

The methods proposed here can be classified under the broad headings:

1. Transformations of parameter space.
2. Sequential minimization.
3. Nested minimization.

The main difficulties in using these methods are that they require the programmer to understand the statistical nature of the model, that they require extra programming for each problem, and that the minimization algorithms must be flexibly written to allow several different minimizations in a single run. The advantages are that convergence is generally rapid, sure and accurate, that initial values do not have to be supplied by the user, that insoluble cases can be detected at an early stage and that an appropriate statistical analysis can be provided. This means that standard library routines can be provided for the more frequently occurring problems, demanding no special skill from the user.

2. Transformation of Parameter Space

In linear estimation problems the concept of orthogonalization is familiar as a simplifying technique that allows parameters to be estimated independently and confidence regions and significance tests to be evaluated. Orthogonalization is a rotation of parameter space which meets objective (c) of Section 1.3. For linear models objective (b) is already satisfied.

In non-linear estimation it is possible to use orthogonalization with good effect, but there is no single linear transformation which will satisfy objectives (b) and (c) over a sufficiently large region to include all plausible initial values. These objectives can sometimes be met by non-linear transformations.

The single-parameter case is an exception in that objective (c) is not relevant, but non-linear transformations can be sought to make the likelihood symmetrical and quadratic in the neighbourhood of the minimum, analogous to the transformations used on data to convert binomial or Poisson variates to approximately normal variates.

Computationally all that is required is that the new parameters should be computable from the old ones, and vice versa. The transformation is not necessarily 1–1; many multicomponent models are parametrized ambiguously, and some convention is required when the old parameters are recovered. Transformations can also clarify what happens to the model when certain parameters become infinite or complex. In particular a parameter with an infinite range can be transformed to one with a finite range, which simplifies the problem of searching parameter space.

2.1. *Stable Parameters*

The choice of transformation is eased if parameters can be found that vary little in the whole region of best-fitting models. Such parameters will simultaneously help

to meet objectives (a), (b) and (c) of Section 1.3. These parameters may be called *stable parameters* because they are little affected by changes in the remaining parameters. They may be found by asking the question:

"What properties are common to all members of the family that fit the data well?"

The answer depends on the goodness of fit of the model. If the fitted values are reasonably close to the observed data then many properties of well-fitting members of the family will vary within a narrow range. This principle may not hold when the model fits very badly, but it is preferable to favour sets of data which do fit the model, and to assume that the suggested transformations will not make the estimation any more difficult for other sets of data. This is best exemplified by returning to Fig. 1.

The three curves A, B and C all have residual sums of squares less than 1 and fit the data well, as would a freehand curve drawn through the points. Therefore all such curves should cut the mean ordinate $x = 3$ at much the same point $y = h$, say, and h is likely to lie in the range (3·5, 4·5) for all such curves. Similarly the area enclosed by the ordinates $x = 1$ and $x = 5$, the x axis and the curve is likely to vary little for all such curves. Hence we expect the following parameters to be stable:

$$h = b(1-r^3), \tag{2}$$

$$a = \int_1^5 b(1-r^x)\,dx = b\left(4-\frac{r^5-r}{\log r}\right) \tag{3}$$

and we can present the results in Table 2.

TABLE 2

Curve	*r*	*b*	*h*	*a*	*R*
A	0·50	4·52	3·96	14·28	0·67
B	0·65	5·35	3·88	14·77	0·27
C	0·80	7·58	3·70	15·02	0·63
D	0·50	7·00	6·13	22·12	21·57
E	0·80	4·00	1·95	7·93	16·09

The effect on the likelihood function when h is used instead of b is illustrated in Fig. 3. The residual sum of squares, as a function of r and h will be called $R'(r, h)$. The contour $R' = 1$ is now almost an ellipse with little interaction between r and h, and because $r = 0$ gives $h = 3{\cdot}6$, a data-based initial value for h is available, namely the mean observed y. The transformation to (r, a) space is very similar but is not illustrated. The limiting case when $r = 1$ is a straight line through the origin and the point $(3, h)$. The lines of best fit are represented in Fig. 3 but not in Fig. 2 because the parameter b becomes infinite as r tends to 1.

2.2. *The choice of stable parameters*

The choice of stable parameters should ideally be such that the transformation will be appropriate whatever data are represented to be fitted to the model. Thus a preliminary descriptive analysis of the supplied data must be made, and parameters derived therefrom. This analysis conveniently forms the basis for automatic initial value estimation.

When fitting distributions the moments can be used in the search for stable parameters. If the expected moments can be expressed in terms of the original parameters then for well-fitting models the expected moments should be close to the observed moments. Thus whereas Pearson's method of moments estimates the original parameters from the observed moments, maximum-likelihood solutions may be sought by

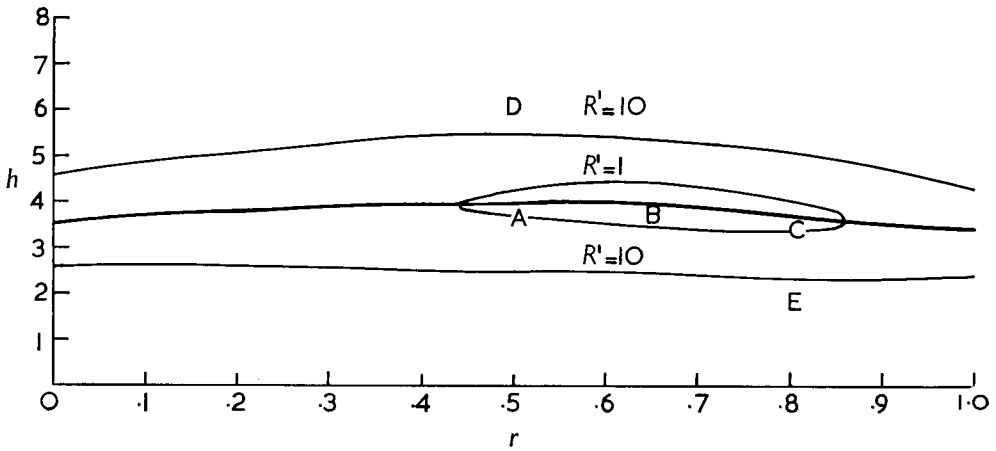


FIG. 3. Contours of the function $R'(r, h)$, where $h = b(1-r^3)$, the intercept of the curve for $x = 3$.

treating the expected moments as parameters and calculating the original parameters therefrom. It is not necessary to transform all the parameters to moment parameters, as the higher moments are less stable than the lower moments. As a simple example consider the negative binomial distribution (Anscombe, 1950), which is often described in terms of its mean m (estimated by the sample mean) and a dispersion parameter k, which is independent of m. The expected variance (which should be stable) is given by the formula

$$v = m+(m^2/k). \tag{4}$$

Hence, when the optimum value of v is determined, the required maximum likelihood estimate of k can be derived from the inverse formula

$$k = m^2/(v-m). \tag{5}$$

This method is particularly effective when k is large, which means that the distribution differs little from the Poisson distribution.

The stable parameters of fitted curves are less easily identified, but intercepts, slopes and definite integrals can all be of use, as shown in the previous section. The parameters of simpler curves fitted to the data or transforms of the data can also be used.

2.3. *Other Transformations*

When stable parameters cannot be identified, or the original parameters expressed in terms of the stable parameters, simpler empirical transformations can be used to good effect, although they are less likely to work well for all possible data. Such a

transformation was found that when fitting the probit plane (Finney, 1952, chapter 7), a relationship between a quantal response and two independent variables of the form

$$E(r_i/n_i) = \Phi(a+bx_{1i}+cx_{2i}), \tag{6}$$

where Φ is a sigmoid function, such as the standard normal integral or the logistic function, and r_i is the number responding out of a random sample of n_i when stimuli x_{1i} and x_{2i} are applied; x_{1i} and x_{2i} in equation (6) should always be expressed as deviations from their means. In this case it was found effective to rewrite the argument of Φ as

$$\theta_1(a_0+b_0x_{1i}+c_0x_{2i})+\theta_2x_{1i}+\theta_3x_{2i}, \tag{7}$$

where a_0, b_0 and c_0 are initial estimates of a, b and c, and the exploration of $(\theta_1, \theta_2, \theta_3,)$ space could begin at $(1, 0, 0)$.

Empirical transformations can be deduced from the study of the likelihood function of a particular problem, but unless they chance to resemble transformations based on stable parameters they are unlikely to have general validity.

3. Sequential Minimization

The parameters of a complicated model can often be estimated from a sequence of simpler models. This is equivalent to invoking principles (d) and (a) of Section 1.3, for by solving an estimation problem in parameter space of few dimensions we are led to good initial estimates for the general problem.

There are many ways in which sequential minimization can operate. With two parameters, one would define some line or curve in the parameter space and minimize the function on that curve. This is equivalent to introducing a constraint in the two-dimensional problem to convert it into a one-dimensional problem. A suitable strategy is to constrain a stable parameter to be constant, with a value estimated from the data by some simple method. For example, in Fig. 2 we can draw the curve corresponding to $h = 4$ shown there as a broken curve. The value $h = 4$ is the observed value of y at $x = 3$. The minimum along this curve corresponds to $r = 0{\cdot}63$, which gives, by equation (2), excellent initial values (0·63, 5·33) for the search in (r, b) space.

The nature of the model may suggest how the search should be organized. For example, when fitting parallel curves to subsets of data it is advisable to fit individual curves to each set and thence to deduce initial values for the full model of parallel curves.

Sequential minimization is particularly powerful when applied to fitting distributions because the low-order moments are very stable. The example of the double normal distribution will now be discussed.

3.1. *The Double Normal Distribution*

The problem of resolving a frequency distribution into two normal components was solved by Karl Pearson (1894) in terms of the sample moments, requiring the roots of a ninth-degree equation. Rao (1948) proposed the maximum-likelihood solution but the practical computations have proved extremely troublesome, and in Rao's numerical example his solution may be significantly improved. Sequential minimization and parameter transformation are effective techniques for this problem.

The method is to fit three different models in order:

Model 1, a single normal distribution,

Model 2, a 50–50 mixture of normal components with equal variances but different means, and

Model 3, two unequal normal components with equal variances.

Initial values of the parameters are obtained at each stage by constraining the expected mean and variance of the combined distribution to be constants, derived from the previous stage. If the components are present in very unequal proportions it is not possible to fit the intermediate model, and this case can be detected when the kurtosis is positive.

If the model to be fitted is

$$f(x) = \alpha N(\mu_1, \sigma^2) + (1-\alpha) N(\mu_2, \sigma^2), \tag{8}$$

where $0 < \alpha < 1$, then the expected moments are as follows:

$$M_1 = \alpha\mu_1 + (1-\alpha)\mu_2, \tag{9}$$

$$M_2 = \sigma^2 + \alpha(1-\alpha)(\mu_1-\mu_2)^2, \tag{10}$$

$$M_3 = \alpha(1-\alpha)(1-2\alpha)(\mu_1-\mu_2)^3, \tag{11}$$

$$M_4 - 3M_2^2 = \alpha(1-\alpha)(1-6\alpha(1-\alpha))(\mu_1-\mu_2)^4. \tag{12}$$

Given a grouped frequency distribution with class boundaries $x_1, \ldots, x_k$, such that n_i observations fall within the interval (x_{i-1}, x_i), $i = 1, \ldots, k+1$, with $x_0 = -\infty$ and $x_{k+1} = +\infty$; then if the distribution function is $F(x)$ the negative log likelihood is

$$L(\boldsymbol{\theta}) = -\sum_{i=1}^{k+1} n_i \log\{F(x_i) - F(x_{i-1})\}, \tag{13}$$

where $\boldsymbol{\theta}$ is the set of parameters defining the model.

The absolute minimum $L_{\min}$ of $L(\boldsymbol{\theta})$ is obtained by equating observed and expected frequencies, thus

$$L_{\min} = N \log N - \sum_{i=1}^{k+1} n_i \log n_i, \tag{14}$$

where

$$N = \sum_{i=1}^{k+1} n_i.$$

This allows the likelihood to be interpreted in terms of the likelihood ratio criterion of Neyman and Pearson (1928), giving a direct test of goodness of fit of each model and the significance of added terms. Hence the estimation can be combined with a series of significance tests.

The three models are as follows:

$$\text{Model 1: } \mu_1 = \mu_2,$$

$$\text{Model 2: } \alpha = 0{\cdot}5, \mu_1 \neq \mu_2,$$

$$\text{Model 3: } \alpha \neq 0{\cdot}5, \mu_1 \neq \mu_2.$$

This method can be extended to fit models with unequal variances or more than two components, but such models tend to suffer from multiplicity of solutions and to introduce more parameters than are justified by the data.

Fitting the general model proceeds in six stages, as follows:

(1) Calculate the sample mean and standard deviation.
(2) Use the optimization routine to fit Model 1 by maximum likelihood. The mean and standard deviation will be slightly different from the sample values because of grouping.
(3) If the kurtosis is positive omit Model 2 and continue from stage (5) with $\alpha = 0{\cdot}15$ if the skewness is negative, and $\alpha = 0{\cdot}85$ if the skewness is positive. These values of α satisfy the inequality of $6\alpha(1-\alpha) < 1$ as required by equation (12) for positive values of $M_4 - 3M_2^2$. Otherwise fit Model 2 by optimizing with respect to σ subject to the expected mean and standard deviation being the values obtained at stage (2). The parameters μ_1 and μ_2 ($\mu_1 < \mu_2$) can then be derived from equations (14).
(4) Fit Model 2 by optimizing with respect to μ_1, μ_2 and σ.
(5) Fit Model 3 with respect to α and σ with the mean and variance constrained as in stage (3), and deriving the parameters μ_1 and μ_2 from equations (9) and (10).
(6) Fit Model 3 by optimizing with respect to μ_1, μ_2, σ and α.

An example of this method is given in Table 3.

TABLE 3

Fitting the double normal distribution

Data and fitted values:

$x <$	n	*Model* 1	*Model* 2	*Model* 3
1	2	4·2	3·0	3·5
2	11	8·0	11·0	11·3
3	27	16·6	25·3	22·0
4	21	27·6	32·8	24·2
5	22	36·4	29·0	23·2
6	36	38·2	29·1	33·8
7	45	31·9	32·8	42·0
8	23	21·2	25·2	29·2
9	11	11·2	10·9	10·6
	4	6·8	2·9	2·2

Fitting process:

Stage	μ_1	μ_2	σ	α	*Likelihood*
1	5·21	—	2·06		11·35
2	5·21	—	2·07		11·34
3	3·64	6·79	1·34		7·79
4	3·36	6·63	1·25		6·67
5	2·99	6·48	1·21	0·36	2·66
6	3·01	6·46	1·20	0·36	2·63

Analysis of χ^2:

Source	χ^2	*d.f.*
Model 1 v. Model 2	9·3	1
Model 2 v. Model 3	8·1	1
Model 3	5·3	5

4. Nested Minimization

This heading includes all methods where the parameters can be arranged in a hierarchy, in which, when particular values can be postulated for parameters of higher rank, optimum values can be estimated for the parameters of lower rank. The simplest situation, which is very common, is that optimum values can be estimated analytically, when it makes little sense to ignore the opportunity to reduce the number of dimensions in which a numerical search is required. The general case involves recursive use of the minimization routine, which requires careful programming, but when there is only one parameter in the highest rank this is not serious.

Analytical minimization is used in curve fitting to estimate the linear parameters when values of the non-linear parameters are assumed. This method was used by Richards (1961) but is frequently overlooked. When particular values of non-linear parameters are assumed the model reduces to linear regression and the linear parameters and function value are directly determined by matrix inversion. Linear parameters defining scales, origins and proportions occur in most curve-fitting applications. A different kind of example is that of Wadley's problem (Finney, 1952) in which a quantal response of an unknown sample size is observed. The unknown size parameter can be fitted as a function of the remaining parameters. Lawley (1967) has applied a similar method to maximum-likelihood factor analysis.

The basic method of analytical minimization is to differentiate the likelihood with respect to the parameters and obtain the standard normal equations, and to eliminate from the likelihood function any parameters conveniently expressed in terms of others.

Nested minimization can be described geometrically as the search of the orthogonal projection onto a sub-set of the parameters of the locus of the minimum over the complementary sub-set. For example, in Fig. 2 the curve ABC is the locus of points for which R is a minimum, given r. The reduced function $\min_b R(r, b)$ is a function of r only and is shown in Fig. 4. In effect the problem is reduced to a one-dimensional search along the curve ABC.

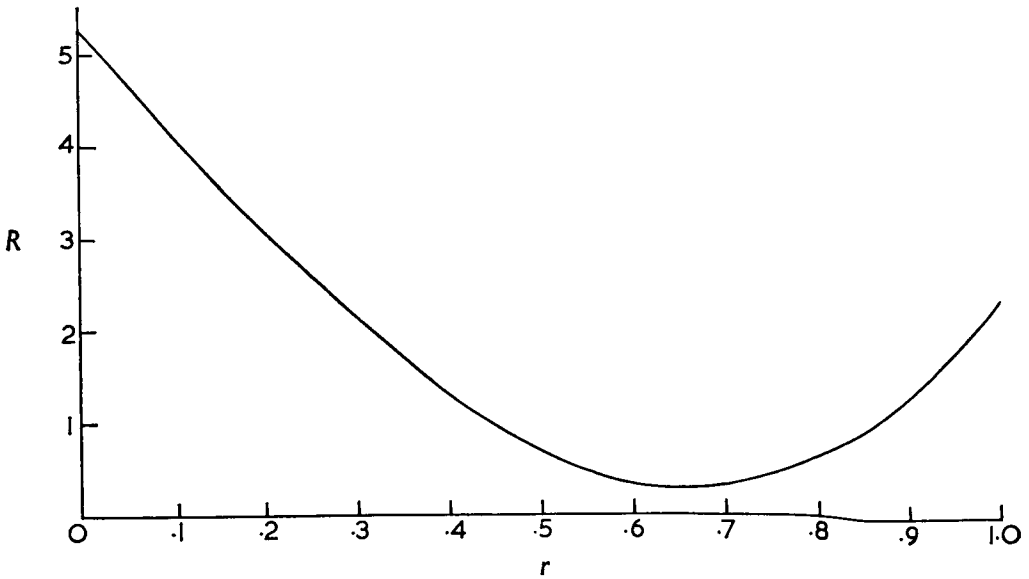


FIG. 4. The function $\min_b R(r, b)$ showing a minimum at $r = 0{\cdot}66$.

4.1. *The Exponential Example*

The example of Fig. 2 works in detail as follows. The residual sum of squares to be minimized is

$$R(r, b) = \sum_{i=1}^{n} \{y_i - b(1 - r^{x_i})\}^2 \tag{15}$$

and the normal equations are

$$b \sum (1 - r^{x_i})^2 = \sum y_i (1 - r^{x_i}) \tag{16}$$

and the less useful equation $\partial R/\partial r = 0$. Then b is eliminated from (15) and (16) to give the function

$$R'(r) = \min_b R(r, b) = \sum y_i^2 - \frac{\{\sum y_i (1 - r^{x_i})\}^2}{\sum (1 - r^{x_i})^2}, \tag{17}$$

which is shown in Fig. 4. The minimum of $R'(r)$ is 0·267 at $r = 0{\cdot}656$ and the corresponding b is 5·397. It happens that $R'(r)$ is nearly a parabola over a wide range of r-values, but this is an accident of scaling, because a change of scale of x to px is equivalent to a transformation from r to r^p. In general, for the best results the non-linear parameters should be transformed so that conditions (b) and (c) of Section 1.3 hold for the simplified function in the restricted parameter space.

4.2. *The HCK Model: a Seven-parameter Model in Bioassay*

A problem solved using nested minimization was a bioassay model fitted to unpublished data of Henry *et al.* (1961). This model will be known as the "HCK model". Two substances are observed to give a sigmoid relationship between a quantitative response (y) and log dose (x), but the range of each curve is unknown as are the location and slope parameters.

The model had seven parameters, as follows:

$$\left.\begin{aligned} &\text{Control group: } y = \theta_1, \\ &\text{Substance 1: } \quad y = \theta_1 + \theta_4 \Phi(\theta_2 + \theta_3 x), \\ &\text{Substance 2: } \quad y = \theta_1 + \theta_7 \Phi(\theta_5 + \theta_6 x), \end{aligned}\right\} \tag{18}$$

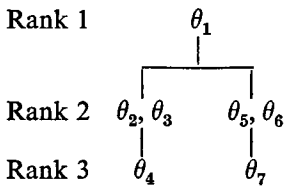
where Φ is the standard normal integral. The data are given in Table 4.

TABLE 4

Data for the HCK model

	x	y
Control	—	87·08
Substance 1	1·59934	98·60
	1·90940	109·22
	2·07733	127·07
	2·31160	145·27
	2·52957	161·83
Substance 2	1·36398	91·13
	1·91840	111·57
	2·09123	114·75
	2·32533	130·68
	2·56949	128·48

Fitting this model proved extremely troublesome to single-pass optimization programs exploring seven-dimensional parameter space, several of the parameters being almost completely correlated. However, the structure of model (18) can be expressed diagrammatically as



This diagram shows that given θ_1 there are two independent sub-problems, the estimation of parameters for substances 1 and 2 separately. Furthermore, θ_4 and θ_7 are simple scaling parameters for which analytical estimates are available. The correlations between θ_2 and θ_3 can be reduced by subtracting the mean, 2·08545 from x for substance 1, and similarly for the correlations between θ_5 and θ_6, by subtracting 2·05369 from x for substance 2. The method therefore was to optimize θ_1, each function evaluation involving separate optimizations for θ_2 and θ_3 and for θ_5 and θ_6. Initial estimates for θ_2 and θ_3 were obtained by storing estimates of $\partial\theta_2/\partial\theta_1$ and $\partial\theta_3/\partial\theta_1$ from previous iterations, since for small adjustments to θ_1 there were almost linear relationships between θ_1 and the optimum values of θ_2, θ_3, θ_5 and θ_6. The three components of the function of θ_1 are shown in Fig. 5 and the results, for Φ as the normal distribution functions, were as follows:

$$\theta_1 = 88{\cdot}09,$$

$$\theta_2 = -4{\cdot}88, \quad \theta_3 = 2{\cdot}20, \quad \theta_4 = 98{\cdot}41$$

$$\theta_5 = -5{\cdot}46, \quad \theta_6 = 2{\cdot}86, \quad \theta_7 = 43{\cdot}57.$$

Residual sum of squares = 55·65.

A rather better fit was obtained using the logistic function (residual sum of squares = 52·92).

The matrix of variances and covariances was computed from the normal equations for all seven parameters given the above results, but is not shown here.

If the minimization routine had not been written recursively, a second *ad hoc* minimization would have sufficed for θ_1.

4.3. *The Use of Nested Minimization*

Nested minimization is usually applied when the optimum in the space of lower rank is unique and is easily determined either by direct calculation or by minimization. For example, if it were applied to Rosenbrock's test function (Rosenbrock, 1960):

$$f(\theta_1, \theta_2) = 100(\theta_2 - \theta_1^2)^2 + (1 - \theta_1)^2, \tag{19}$$

it would be necessary to rank θ_1 above θ_2 because solutions of θ_1 given θ_2 are not unique.

Nested minimization arises naturally when a model is applied to several sets of data simultaneously, as in the HCK model or in fitting parallel curves. There are *general* parameters applying to all sets of data, and *specific* parameters applying only to a particular set. For example, in the linear model of parallel linear regression the

common slope is a general parameter where the individual intercepts are specific parameters.

The low dimensional functions obtained by nested minimization can be more easily studied than the full likelihood function, and the difficulties associated with some models can be more readily understood. Contours of values of subsidiary parameters can also be constructed and superimposed on the likelihood contours.

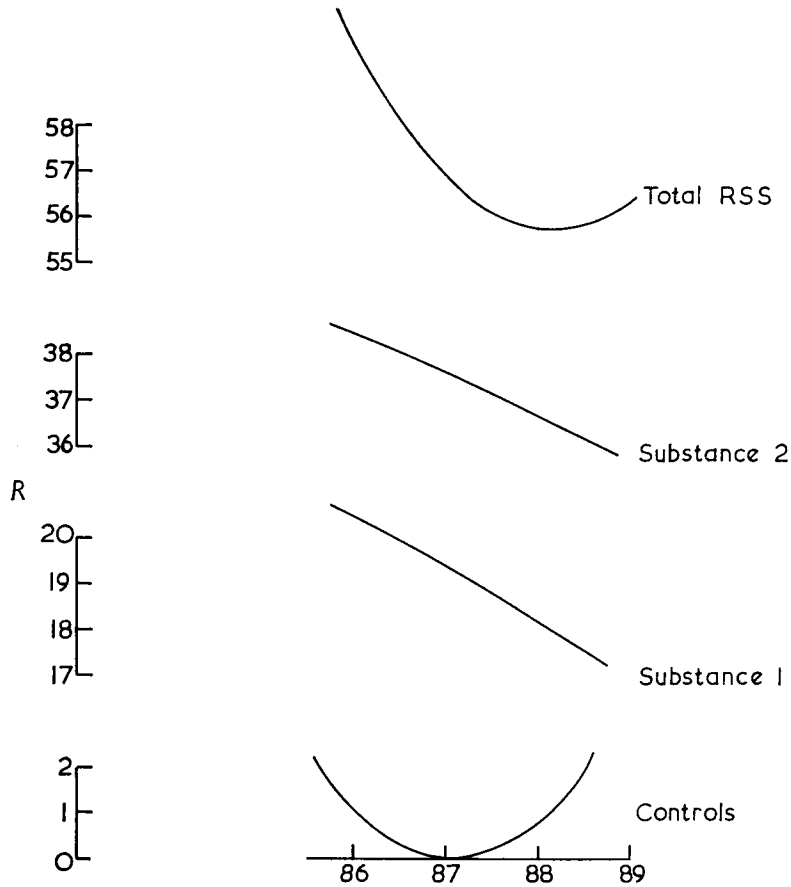


FIG. 5. The HCK model. The residual sum of squares R as a function of θ_1, and the three independent components of R.

5. COMPUTING CONSIDERATIONS

The methods discussed above require slightly more elaborate programming than is found in standard library programs for optimization, and the following questions arise:

(1) Is the increase in computing efficiency worth the extra trouble when computers are becoming so fast anyway?

(2) Can the methods be generalized so that the best strategy can be selected automatically?

(3) How should optimization programs be written to make these methods simple to implement?

5.1. *Comparison with Simple Minimization*

Minimization methods are often compared by quoting the number of function evaluations required to solve standard problems. This is not exactly a fair comparison in the present case because the functions to be minimized are not comparable although they do not in general involve much more computing than the straightforward likelihood. (Transformed functions take much the same time, sequential and nested functions may take more or less time according to circumstances.) The comparison is also difficult because the objectives of problem analysis include the provision of automatic initial values, recognition of insoluble cases and diminution of any tendency to diverge. Transformations are designed to favour minimization methods that are quadratically convergent.

The examples in this paper have been tested using a quadratically convergent minimization algorithm based on the Taylor series expansion of a general function. This method did not work well when the problems were badly defined, but was satisfactory when transformations were used. To illustrate the effect of reducing dimensionality by analytical elimination of parameters, Table 5 gives some typical

TABLE 5

Effect of analytical elimination of parameters

Model	No. of parameters		No. of function calls		Model	Standard method	Reference
	Full	Non-linear	Standard	Reduced			
1	3	1	191	10	$y = a+br^x$	Simplex	—
2	4	1	452	10	$y = a+(b+cx)\,r^x$	Simplex	—
3	5	2	503	68	$y = a+br^x+cs^y$	Simplex	—
4	3	2	228†	15	Rational function	Fletcher and Powell	Chambers
5	6	1	120†	10	Quantal responses	Fletcher and Powell	Chambers
6	4	2	400	54	Enzyme reactions	Rosenbrock	Kowalik

† Unsatisfactory convergence.

comparisons, the first three examples being exponential curves fitted by Nelder and Mead's simplex algorithm, examples 4 and 5 being taken from Chambers (1969) (ex. 2 and 3) and example 6 from Kowalik and Osborne's problem (v). Examples 1, 2, 4 and 5 can be solved on a desk calculator.

Sequential and nested minimization, with transformations, made problems such as the double normal distribution and the HCK model quite straightforward whereas they would not converge at all when tackled directly. But the really important point is that these methods are suitable for library programs for standard models, in which the user does not have to provide initial values or operational quantities for the minimization routine. Ideally the minimization process should be concealed from the user, unless it fails to converge, as it is no more important to him than the details of a matrix inversion.

5.2. *Automatic Strategy Selection*

There seems to be no obvious way in which numerical analysis can replace statistical insight as a means of controlling the likelihood function. Non-linear transformations can only be built up by extensive exploration of the function, whereas statistical insight enables useful transformations to be anticipated. Automatic nested minimization is possible but tends to be expensive in terms of numbers of function evaluations.

5.3. *Writing Optimization Routines*

The methods described are not suitable for use with minimization routines that require derivatives or assume that the function is a sum of squares. The derivatives are often difficult to compute directly. The minimization routine is a sub-routine which is called upon to minimize several different functions in sequence, with parameters fixed or varied as required. If the sub-routine cannot be written recursively, a separate routine can be written for the outer minimization that effectively enables recursions to be used.

Numerical accuracy is important, and double length working may be required, especially when analytical elimination is used, and in nested minimization the convergence must be strict enough to provide the outer minimizations with a smooth function. The minimizer should be able to recognize situations in which the function to be minimized is effectively constant, which can occur when analytical elimination is used and the current estimates are too inaccurate. The reason for this is that non-linear terms such as exponentials can become effectively constant for all the data values, and there is therefore no information available to the minimizer. In such cases the program must appeal to the user to provide fresh starting values, or to check for gross errors in the data, or to re-scale the data. Sometimes the reason for convergence failure can be seen, and the model modified automatically if the first attempt fails. For example, a convex exponential curve will not fit data that is basically concave, but the minimization will then end with the exponential parameter taking its limiting value and the process can be restarted with a change of sign.

6. Conclusions

I have described how individual problems can be solved by adopting a flexible attitude to the use of minimization routines, by using the clues in the descriptive statistics of the data and in the analytical and statistical properties of the model. There is not necessarily one ideal strategy for all models, and several unrelated techniques can give numerically satisfactory results, as shown by the various methods used on the model of Fig. 1.

The benefits of special programming as contrasted with straightforward use of the original model can be summarized as follows:

(1) Use of appropriate parameter transformations, sequential and nested minimization can greatly shorten computing time.
(2) The estimates of the parameters tend to be more accurate because parabolic approximations are justified over a wider region and nested parameters are estimated exactly.
(3) Convergence is more certain, and reasons for failure can be readily studied.
(4) In programs for standard problems the mechanics of minimization may be suppressed from the results, giving the user the same kind of output he would expect from a linear regression program. In particular, the user does not have to supply quantities needed to make the minimizer work effectively.

(5) Satisfactory confidence regions can be obtained for the parameters and fitted values of the data if the transformed likelihood function is approximately parabolic.

(6) Sequential minimization combines a method of solution with a statistical analysis of successive models.

The disadvantages are for the user who wants an answer with the minimum of programming effort. Yet the appropriate devices are not difficult to recognize and to incorporate in a program, and a little forethought may save much time later. There are many reports in print of problems that have proved "difficult" for approved minimization routines, and the reason has usually been lack of attention to the possibilities described above. A further difficulty is that not all widely available minimization programs make it easy for the user to adopt these special devices, although it is to be hoped that this will not always be so.

The ideas described here have been incorporated in a library program, MLP, written for the Orion computer at Rothamsted. The program, written in machine code, can be used either as a standard set of routines for the more frequently occurring models, or as a sub-routine package for new models. Details of these standard routines will be published separately.

References

ANSCOMBE, F. J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, **37**, 358–382.

BOX, M. J., DAVIES, D. and SWANN, W. H. (1969). Non-linear optimisation techniques. *I.C.I. Monograph No. 5*. Edinburgh: Oliver & Boyd.

CHAMBERS, J. M. (1969). FIT, a new approach to fitting a model to data: IV. User's model. (To be published.)

FINNEY, D. J. (1952). *Probit Analysis*, 2nd ed. Cambridge: University Press.

FLETCHER, R. and POWELL, M. J. D. (1963). A rapidly convergent descent method for minimization. *Computer J.*, **6**, 163–168.

HENRY, KATHLEEN M., CORMACK, R. M. and KOSTERLITZ, H. W. (1961). The determination of the nutritive value of a protein by its effect on liver nitrogen in rats. *Brit. J. Nutr.*, **15**, 199–212.

KOWALIK, J. and OSBORNE, M. R. (1968). *Methods for Unconstrained Optimization Problems*. New York: Elsevier.

LAWLEY, D. N. (1967). Some new results in maximum likelihood factor analysis. *Proc. Roy. Soc., Edinburgh*, **67**, 256–264.

MARQUARDT, D. W. (1963). An algorithm for least squares estimation of non-linear parameters. *J. Soc. Indust. Appl. Math.*, **11**, 431–441.

NELDER, J. A. (1966). Inverse polynomials, a useful group of multi-factor response functions. *Biometrics*, **22**, 128–141.

NELDER, J. A. and MEAD, R. (1965). A simplex method for function minimisation. *Computer J.*, **7**, 308–313.

NEYMAN, J. and PEARSON, E. S. (1928). On the use and interpretation of certain test criteria for the purposes of statistical inference. *Biometrika*, **20**, 175–240.

PEARSON, K. (1894). Contributions to the mathematical theory of evolution. *Phil. Trans. Roy. Soc.*, **185**, 71–110.

POWELL, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer J.*, **7**, 155–162.

—— (1965). A method for minimizing a sum of squares of non-linear functions without calculating derivations. *Computer J.*, **7**, 303–307.

RAO, C. R. (1948). The utilisation of multiple measurements in problems of biological classification. *J. R. Statist. Soc.* B, **10**, 159–203.

RICHARDS, F. S. G. (1961). A method of maximum likelihood estimation. *J. R. Statist. Soc.* B, **23**, 469–476.

ROSENBROCK, H. H. (1960). An automatic method for finding the greatest or least value of a function. *Computer J.*, **3**, 175–184.