Nonlinear Parameter Estimation

YONATHAN BARD

Nonlinear Parameter Estimation

YONATHAN BARD
International Business Machines Corporation
Cambridge, Massachusetts

ACADEMIC PRESS New York and London 1974
A Subsidiary of Harcourt Brace Jovanovich, Publishers

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1

Library of Congress Cataloging in Publication Data

Bard, Yonathan.
Nonlinear parameter estimation.

Bibliography: p.
1. Estimation theory. I. Title.
QA276.8.B37 519.5'4 72-13616
ISBN 0-12-078250-2

AMS(MOS) 1970 Subject classifications: 62F10, 62J05, 90C30

PRINTED IN THE UNITED STATES OF AMERICA

Preface

This book is intended primarily for use by the scientist or engineer who is concerned with fitting mathematical models to numerical data, and for use in courses on data analysis which deal with that subject. Such fitting is frequently done by the method of least squares, with no regard paid to previous knowledge concerning the values of the parameters (coefficients), nor to the statistical nature of the measurement errors. In Chapters II–IV we show how the problem can be formulated so as to take all these factors into account. In Chapters V–VI we discuss the computational methods used to solve the problem, once its formulation has been completed. Chapter VII is devoted to the question of what conclusions can be drawn, after the estimates have been computed, concerning the validity of the estimates, or of the model which has been fitted. In Chapter VIII we discuss the important special case of models which are stated in the form of differential equations. Other special problems are treated in Chapter IX. Finally, in Chapter X we suggest methods for planning the experiments in such a way that the data will shed the greatest possible light on the model and its parameters. We cannot stress too strongly the point that if data are to be gathered for the purpose of establishing a mathematical model, then the experiments should be designed with this purpose in mind. Hence the importance of Chapter X.

A practical, rather than theoretical point of view has been taken throughout this book. We describe computational algorithms which have performed well on a variety of problems, even if their convergence has not been proven, and even if they have failed on some other problems. We have as yet no foolproof, efficient methods for solving nonlinear problems; hence we cannot afford to throw away useful tools just because they are not perfect.

The presentation uses matrix algebra and probability theory on a very elementary level. Reviews of the needed concepts and proofs of some important theorems will be found in the appendixes. Some supplementary material has been included in the form of problems at the ends of chapters. Problems requiring actual computation have not been included; the reader is likely to have his own data to compute with, and additional data may be found in many of the cited references. Several numerical problems have, however, been worked out in great detail in separate sections at the ends of Chapters V–IX for the purpose of illustrating the methods discussed in those chapters.

The author is deeply indebted to the IBM Corporation, and in particular to the managements of the New York and Cambridge Scientific Centers, who have supported the writing of this book and provided all the necessary resources. The author is also grateful to Professor L. Lapidus of Princeton University, and to his colleagues J. G. Greenstadt, P. G. Comba, H. Eisenpress, K. Spielberg, and P. Backer, for many helpful discussions, and for reviewing portions of the manuscript.

Contents

Chapter

1 Introduction

1-1. Curve Fitting

A scientist who has compiled tables of data wishes to reduce them to a more convenient and comprehensible form. He accomplishes this by representing the data in graphical or functional form. In the first case, he plots his data points, and then draws some curve through them. In the second case, he selects a class of functions, and chooses from this class the one that best fits his data. This is called *curve fitting*.

In the simplest case, the data consist of values $y_1, y_2, \ldots, y_n$ of a dependent variable y measured for various values $x_1, x_2, \ldots, x_n$ of an independent variable x. A frequently chosen class of functions is the set of all polynomials of order not exceeding m

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_m x^m \tag{1-1-1}$$

The values of the *parameters* $\theta_0, \theta_1, \ldots, \theta_m$ are chosen so as to get the best possible fit to the data. The most commonly used technique for accomplishing this is the least squares method, in which those values of the θ_i are selected which minimize the sum of squares of the *residuals*, i.e.,

$$S = \sum_{\mu=1}^{n} \left(y_\mu - \sum_{\alpha=0}^{m} \theta_\alpha x_\mu^\alpha \right)^2 \tag{1-1-2}$$

Curve fitting procedures are characterized by two degrees of arbitrariness. First, the class of functions used is arbitrary, being dictated only to a minor extent by the physical nature of the process from which the data came. Second, the best fit criterion is arbitrary, being independent of statistical considerations. This arbitrariness can be exploited to make the fitting job easy. Choosing equations which, like Eq. (1),‡ are linear functions of the parameters; using orthogonal or Fourier polynomials (in place of ordinary

‡ This reference is to the first equation of the current section, i.e., Eq. (1-1-1).

polynomials) as the functions to fit; employing the least squares criterion—all these contribute to making the computation of the parameters a mathematically easy job. On the other hand, due to their arbitrary nature, the equations that we get are useful only for summarizing the data and for interpolating between tabulated values. They cannot be used to extrapolate, i.e., to predict the outcome of experiments removed from the region of already available data. Also, the equations and the parameters occurring in them shed little insight on the nature of the process being measured, except to answer such questions as to whether variable x has an influence on variable y.

Curve fitting techniques have widespread applications in situations that go far beyond the simple y vs. x table. An example is the identification of dynamic systems by means of rational transfer functions or Volterra series. Most multiple linear regression, analysis of variance, and econometric time-series problems are also of a curve fitting nature, since the equations used are not derived from "laws of nature." In most of these applications, however, assumptions are made concerning the statistical behavior of the errors, thereby elevating them at least partly to the status of estimation problems as discussed in Section 1-3.

1-2. Model Fitting

Often the scientist is, to a certain extent, familiar with the laws which govern the behavior of the physical system under observation. He can then derive equations describing the relationships among the observed quantities. For instance, the fraction y of a radioactive isotope remaining x seconds after the isotope's formation is given by

$$y = e^{-\theta x} \tag{1-2-1}$$

where the parameter θ is a physical constant proportional to the instantaneous rate of decay of the isotope. The magnitude of θ is unknown, but we wish to assign to it a value which makes Eq. (1) fit the data (y_1, x_1), (y_2, x_2), ..., (y_n, x_n) as well as possible, e.g., by the least squares criterion.

An equation such as Eq. (1) which is derived from theoretical considerations is called a *model*, and the procedure just described constitutes *model fitting*. In principle, model fitting is not much different from curve fitting, except that we can no longer guide the selection of a functional form by considerations of computational convenience. For instance, Eq. (1) is not a linear function of the parameter, and because of this the computation of the "best fit" is more difficult than the computation of the θ_i in Eq. (1-1-1).

1-3. Estimation

A new consideration arises in model fitting that does not exist in curve fitting. The parameters occurring in a model, e.g., θ in Eq. (1-2-1), usually represent quantities that have physical significance. If the model is a correct one, then it is meaningful to ask what is the true value of θ in nature. Because of the generally imprecise nature of measurements we can never hope to determine the true values with absolute certainty. Also, due to the random nature of the errors in measurements, the value of θ that best fits one series of measurements differs from the value that fits another series, even though both series are performed on the same isotope. However, we can look for procedures to obtain values of the parameters that not only fit the data well, but also come on the average fairly close to the true values, and do not vary excessively from one set of experiments to the next. The process of determining parameter values with these statistical considerations in mind is termed *model estimation*.

The classical problem of *statistical estimation* differs somewhat from the model estimation problem that we have just defined. The statistician observes a sequence of values ("realizations") that a random variable assumes. For instance, he may obtain a sequence of numbers such as 1, 5, 6, 3, . . . denoting successive throws of a die. The statistician assumes a "model" in the form of a probability distribution which may depend on some unknown parameters. In our case, the statistician who suspects the die may be loaded assigns probabilities $[\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, 1 - \sum_{i=1}^{5} \theta_i]$ to the six possible outcomes of a throw. He then attempts to estimate the θ_i from the observed values of the random variable. Here he will probably use the estimate

$$\theta_i = n_i / \sum_{j=1}^{6} n_j \tag{1-3-1}$$

where n_i is the number of throws on which the number i showed up ($i = 1, 2, \ldots, 6$).

As a further example, the observed value of the random variable may be the height h of adults in a community. If we assume that this variable has normal (Gaussian) distribution with mean h_0 and standard deviation σ, then the probability density function is given by

$$p(h) = [1/(2\pi)^{1/2}\sigma] \exp[-(1/2\sigma^2)(h - h_0)^2] \tag{1-3-2}$$

If we measure the heights $h_1, h_2, \ldots, h_n$ of n randomly chosen individuals from the community, we form the usual estimates:

$$h_0 = (1/n) \sum_{\mu=1}^{n} h_\mu \tag{1-3-3}$$

$$\sigma^2 = [1/(n-1)] \sum_{\mu=1}^{n} (h_\mu - h_0)^2 \tag{1-3-4}$$

The model estimation problem can be embedded in the statistical estimation problem in the following way: It is reasonable to suppose that the outcome y of a measurement taken at time x_μ (we shall phrase our discussion in terms of the radioactive decay model of Section 1-2) is a random variable whose mean value is given by Eq. (1-2-1) as $\exp(-\theta x_\mu)$. If many measurements were to be taken at the same x_μ we would discover that the observed values y_μ fluctuate around their mean value with standard deviation σ. Suppose these fluctuations have a normal probability distribution. The probability density function for y_μ would then have the form similar to Eq. (2)

$$p(y_\mu) = [1/(2\pi)^{1/2}\sigma] \exp\{-(1/2\sigma^2)[y_\mu - \exp(-\theta x_\mu)]^2\} \tag{1-3-5}$$

In fact we only take one measurement at any specific x_μ. What we have are realizations $y_1, y_2, \ldots, y_n$, each of a different random variable whose distribution depends on the parameter x_μ which varies from one variable to the next, and on some other parameters (θ, σ) which are common to all these distributions. The *parameter estimation* problem which is the primary concern of this book is the problem of estimating these common parameters.

At first glance, the parameter estimation problem appears more general than the classical statistical estimation problem, since in the latter all samples are taken from the same distribution. The distinction between the two problems disappears if we choose to regard all the data as being a single multivariate sample from the joint distribution of all the observations made in the course of the series of experiments. It follows that many of the statistical estimation methods can be applied to our parameter estimation problems. The single sample point of view is, however, rather awkward when one examines, say, the asymptotic properties of these estimates (see Chapter III for definitions) since it requires that the entire set of experiments be repeated over and over again.

Parameter estimation techniques may be applied as computational tools to pure curve fitting problems. One must remember, however, that the statistical properties of these estimates (e.g., those described in Chapters III and VII) sometimes lose their meaning in the curve fitting context.

Clearly, parameter estimation is a more difficult operation than curve fitting, calling for more sophisticated analysis and more extensive computation. The effort is worthwhile since a well established model and precisely estimated physical parameters are much more versatile tools, both for illuminating the present situation and for prediction in new situations, than arbitrarily fitted curves can ever be. To bring home this point, one need only observe that a physical parameter estimated from one model can always be used in another model to which it is relevant. For instance, the viscosity of a

liquid estimated from viscometer data can be used to predict the required pumping load for a piping system being designed.

There are other mathematical problems which may be solved by means of parameter estimation or curve fitting techniques. These techniques may be regarded as attempts to solve (as best one can) an overdetermined (more equations than unknowns) system of simultaneous equations. Solving a system of n equations in n unknowns may, therefore, be regarded as fitting to n data points a model involving n unknown parameters. Two-point boundary value problems in ordinary differential equations may be treated as models in which the known terminal conditions are the data, and the missing initial conditions are the unknown parameters. Some optimal control problems may be solved by regarding the control actions as unknown parameters, and the desired trajectory of the system as the data to be fitted. Similarly, some engineering design problems may be posed as requiring parameter values which induce the systems to meet prescribed conditions as closely as possible.

1-4. Linearity

To understand what we mean by the term "nonlinear estimation" we must first make the following definitions: An expression is said to be *linear* in a set of variables $\phi_1, \phi_2, \ldots, \phi_n$ if it has the form $a_0 + \sum_{i=1}^{n} a_i \phi_i$, where the coefficients a_i $(i = 0, 1, \ldots, n)$ are not functions of the ϕ_i. An expression is *quadratic* in the ϕ_i if it has the form $a_0 + \sum_{i=1}^{n} a_i \phi_i + \sum_{i,j=1}^{n} b_{ij} \phi_i \phi_j$, again with all coefficients not depending on the ϕ_i. If we differentiate a quadratic expression with respect to one of the ϕ_i, we obtain a linear expression.

Linear estimation problems are ones in which the model equations are linear expressions in the unknown parameters, e.g., Eq. (1-1-1). When the model equations are not linear, as in Eq. (1-2-1), we speak of *nonlinear estimation*. However even some apparently linear problems are essentially nonlinear. This is so because in order to estimate the parameters we usually minimize some function, such as the sum of squares of residuals. To find the minimum, we equate the derivatives of the function to zero and solve for the values of the parameters. Now when the model equations are linear, the sum of squares function is quadratic, and the derivatives are again linear. The estimates are obtained, therefore, by solving a set of simultaneous linear equations, and all is well. But if some other functions which are not quadratic are chosen to be minimized, then the equations to be solved are no longer linear, even when the model equations are linear. Such problems should also be regarded as nonlinear estimation problems. Examples of such problems are given in Sections 4-8–4-9.

1-5. Point and Interval Estimation

There exist many methods (e.g., least squares) which calculate specific numbers representing estimates for the parameter values. Such numbers are called *point estimates*. A point estimate for the parameters θ, σ appearing in Eq. (1-3-5) may take the form

$$\theta^* = 4, \qquad \sigma^* = 0.1 \tag{1-5-1}$$

A point estimate standing alone is not very satisfactory. Random errors are present in all measurements, and no mathematical model accounts for all facets of a physical situation. Therefore we cannot hope to obtain point estimates exactly equal to the true values of the parameters (if such exist). Nor can we expect point estimates calculated from different data samples to be equal, even if the samples were obtained under similar conditions. Therefore we need to augment the point estimate with some information on its variability. For instance, in place of Eq. (1) we wish to have a statement such as

$$\theta^* = 4 \pm 0.2, \qquad \sigma^* = 0.1 \pm 0.02 \tag{1-5-2}$$

The numbers 0.2 and 0.02 are meant to represent the standard deviations of the variability of the estimates for θ, σ.

The information contained in Eq. (2) may be translated into a statement of the type‡ "We are 75% sure that θ is between 3.6 and 4.4, and we are 75% sure that σ is between 0.06 and 0.14." This statement constitutes an *interval estimate* for our parameters.

Interval estimates can be computed directly, without first calculating point estimates and their variability. In fact, many statisticians prefer interval estimates, because they feel one is not justified in picking out one specific preferred value to be used as a point estimate. We feel, however, that the needs of the scientist or engineer are best served by point estimates with measures of their reliability, so we will not discuss any direct interval estimation procedures. The calculation of interval estimates (called *confidence intervals* in this context) from point estimates is discussed in Sections 7-9–7-10.

1-6. Historical Background

Legendre (1805) was the first to suggest in print the use of the least squares criterion for estimating coefficients in linear curve fitting. Gauss (1809) laid the statistical foundation for parameter estimation by showing that least squares estimates maximized the probability density for a normal (Gaussian)

‡ The statement is derived from Eq. (2) using the Bienaymé–Chebyshev inequality with $k = 2$. See Eq. (7-9-11).

distribution of errors. In this, Gauss anticipated the maximum likelihood method. Gauss and his contemporaries seemed to prefer, however, purely heuristic justifications for the least squares method. Further work in the 19th and early 20th centuries, by Gauss himself, Cauchy, Bienaymé, Chebyshev, Gram, Schmidt, and others‡ concentrated on computational aspects of linear least squares curve fitting, including the introduction of orthogonal polynomials.

The development of statistical estimation methods received its impetus from the work of Karl Pearson around the turn of the century and R. A. Fisher in the 1920s and 1930s. The latter revived the maximum likelihood method and studied estimator properties such as consistency, efficiency, and sufficiency [see the collection of Fisher's (1950) papers]. The development of decision theory by Wald and others has, in the post-World War II years, introduced a new basis for selecting estimation criteria. The practical impact of these methods in the area of nonlinear parameter estimation has so far been slight, except for causing increased awareness of the uses of prior distributions.

The first modern applications of statistical estimation theory to model estimation were made in the field of econometrics by Koopmans and others, starting in the 1930s. Their work is summarized in the Cowles Commission Reports (Hood and Koopmans, 1953). The main contributions to the application of statistical techniques in the construction and estimation of mathematical models in the physical sciences have come from professor G. E. P. Box and his coworkers at Princeton University and the University of Wisconsin.

The computation of estimates for nonlinear models usually requires finding the maximum or minimum of a nonlinear function. Computational methods bearing the names of Newton, Gauss, and Cauchy have been known for a long time, but their extensive application to practical problems had to await the arrival of the electronic computer. The first general purpose computer program for solving nonlinear least squares problems was written by Booth and Peterson (1958) in collaboration with Box. The program employed a modified Gauss method. It has since been followed by many other programs, some more general in nature and some dealing with more specific estimation problems. A list of such programs can be found in Appendix G.

1-7. Notation

Matrix and vector notation are used throughout this book.
A boldface capital letter denotes a matrix: $\mathbf{A}$, $\boldsymbol{\Gamma}$.
A boldface lower case letter denotes a column vector: $\mathbf{a}$, $\boldsymbol{\gamma}$.

‡ References to this work, along with a more detailed historical survey are given by Seal (1967).

The (i,j) element, appearing in the ith row and jth column of $\mathbf{A}$ is denoted A_{ij} or $[\mathbf{A}]_{ij}$.

The ith element of $\mathbf{a}$ is denoted a_i or $[\mathbf{a}]_i$.

$\mathbf{A}_\mu$ is the μth in a sequence of matrices $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \ldots$. The (i,j) element of $\mathbf{A}_\mu$ is denoted $A_{\mu ij}$ or $[\mathbf{A}_\mu]_{ij}$. Analogously for vectors.

$\mathbf{A}^T$ is the transpose of $\mathbf{A}$, i.e., $[\mathbf{A}^T]_{ij} = [\mathbf{A}]_{ji}$.

$\mathbf{a}^T$ is the row vector with the same elements as $\mathbf{a}$.

$\mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$ if such exists.

$\mathbf{A}^+$ is the pseudoinverse of $\mathbf{A}$.

$\det(\mathbf{A})$ is the determinant of $\mathbf{A}$.

$\mathrm{Tr}(\mathbf{A}) = \sum_i A_{ii}$ is the trace of $\mathbf{A}$.

$\mathbf{A}$ is said to be $m \times n$ if it has m rows and n columns. A column vector is $m \times 1$ and a row vector $1 \times n$.

$\mathbf{I}$ is the identity matrix, i.e.,

$$I_{ij} = \delta_{ij} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

$\mathbf{I}_m$ is the $m \times m$ identity matrix.

$\mathbf{A} = \mathrm{diag}(\mathbf{a})$ means that $\mathbf{A}$ is a matrix with elements $A_{ij} = a_i \delta_{ij}$.

Suppose α is a function of the vectors $\mathbf{a}$ and $\mathbf{b}$ and the matrix $\mathbf{A}$. Then:

$$\begin{aligned} \partial\alpha/\partial\mathbf{a} &\text{ is the column vector } [\partial\alpha/\partial\mathbf{a}]_i = \partial\alpha/\partial a_i \\ \partial\alpha/\partial\mathbf{A} &\text{ is the matrix } [\partial\alpha/\partial\mathbf{A}]_{ij} = \partial\alpha/\partial A_{ij} \\ \partial^2\alpha/\partial\mathbf{a}\,\partial\mathbf{b} &\text{ is the matrix } [\partial^2\alpha/\partial\mathbf{a}\,\partial\mathbf{b}]_{ij} = \partial^2\alpha/\partial a_i\,\partial b_j \end{aligned}$$

Suppose $\mathbf{a}$ is a vector function of the scalar β and the vector $\mathbf{b}$. Then:

$$\begin{aligned} \partial\mathbf{a}/\partial\beta &\text{ is the column vector } [\partial\mathbf{a}/\partial\beta]_i = \partial a_i/\partial\beta \\ \partial\mathbf{a}/\partial\mathbf{b} &\text{ is the matrix } [\partial\mathbf{a}/\partial\mathbf{b}]_{ij} = \partial a_i/\partial b_j \end{aligned}$$

Suppose $\mathbf{A}$ is a matrix function of the scalar α. Then:

$$\partial\mathbf{A}/\partial\alpha \text{ is the matrix } [\partial\mathbf{A}/\partial\alpha]_{ij} = \partial A_{ij}/\partial\alpha$$

Derivatives of matrices with respect to vectors and matrices, or of vectors with respect to matrices, give rise to arrays with more than two dimensions. Rules for differentiating vector and matrix expressions are given in Section A-2 Appendix A.

We also make use of some notation associated with probability concepts.

$\Pr(A)$ is the probability of event A.