CHAPTER 1

INTRODUCTION

The purpose of this book is to introduce methods for the solution of a general type of parameter estimation problem often met in applied science and engineering. In these disciplines, observations are usually not the quantities to be measured themselves but are related to these quantities. Throughout the book, it will be assumed that this relation is a known function and that the quantities to be measured are parameters of this function. Thus, parameter estimation is the computation of numerical values for the parameters from the available observations. For example, if observations have been made of the radioactive decay of a compound, the function is a multiexponential decay function of time. The parameters are the amplitudes and the decay constants of the exponential components. The parameter estimation problem is here: computing the amplitudes and the decay constants from the observations.

Applied scientists and engineers agree that most observations contain errors. Clearly, the description of observations as function values for known values of the independent variable, called measurement points, is incomplete. If a particular experiment is repeated under the same conditions, the resulting sets of observations will, as a rule, differ from experiment to experiment. These unpredictable fluctuations of the observations are sometimes called nonsystematic errors. An effective way of describing the fluctuations is by means of statistics. This implies that observations are modeled as stochastic variables. The function values are the expectations of these stochastic observations. The fluctuation is the deviation of an observation from its expectation. The function of the observations used to compute the parameters is called estimator. Since the observations are stochastic variables, so is the estimator. Therefore, the estimator has an expectation and a standard deviation.

One particular outcome of the estimator is called estimate. An estimate is a number. The standard deviation of the estimator defines and quantifies its precision. It is a measure of the magnitude of the nonsystematic error of the estimator caused by the fluctuations of the observations. The estimator is said to be more precise as its standard deviation is smaller. The deviation of the expectation of the estimator from the hypothetical true value of the parameter is called the bias of the estimator. The bias of the estimator defines and quantifies its accuracy. It is a measure of the systematic error of the estimator as a result of the fluctuations of the observations. The estimator is said to be more accurate as its bias is absolutely smaller. If the fluctuations of the observations would be absent, the estimator would produce exact values for the parameters.

A completely different source of error are modeling errors. If the expression for the expectations of the observations and the function used by the experimenter are different parametric families of functions, this will lead to systematic errors in the estimated parameters. For example, if the expectations of observations made are the sum of a multiexponential function and a constant representing background radiation while the function used by the experimenter is only the multiexponential function, the parameters of a wrong function will be estimated. Modeling errors are not to be considered systematic errors in the observations. They are caused by the experimenter. On the other hand, if a background is routinely included in the model although it is known that there is no background present, the model is correct but the background parameter is equal to zero and superfluous. This is an example of overparameterization. Overparameterization increases the standard deviation of the estimates of the remaining parameters, which, in our example, are the amplitudes and the decay constants.

The terminology thus introduced enables us to give the following overview of the book.

In Chapter 2, we discuss parametric models of observations. After a short introduction in Section 2.1, this discussion starts in Section 2.2 with deterministic (that is, nonstatistical) parametric models. This is followed, in Section 2.3, by an analysis of traditional parameter estimation methods which are based on the assumption that such deterministic observations really exist. This analysis reveals the need for the statistical parametric models introduced in Section 2.4. In the same section, the term expectation model is introduced for the parametric function that describes the expectations of the observations.

Describing observations as stochastic variables implies that they are defined by probability (density) functions. These are the subject of Chapter 3. After a short introduction in Section 3.1, we define in Section 3.2 some preliminary notions such as the covariance matrix of a set of stochastic variables and the Fisher score vector. Then, in the Sections 3.3–3.5, three important joint distributions of sets of observations are defined and discussed: the multivariate normal, the Poisson, and the multinomial distribution. These belong to an important general class called exponential families of distributions. This class is introduced in Section 3.6. Exponential families will be used throughout the book and will lead to considerable generalizations and simplifications. Statistical properties of the Fisher score vector are discussed in Section 3.7. In Section 3.8, complex stochastic variables are introduced. These are important in a number of disciplines in applied science dealing with complex parameters or complex observations. As an important example, the joint normal distribution of real and complex stochastic variables is discussed in Section 3.9.

Accuracy and precision of parameter estimators are the subjects of Chapter 4. It is essential to know to what extent observations reduce the uncertainty about the hypothetical true value of a parameter. If we make suitable assumptions about the distribution of the observations, this reduction of uncertainty can be quantified using the concept Fisher information. After a short introduction in Section 4.1 and a review of relevant estimation terminology and properties of covariance matrices in Sections 4.2 and 4.3, we introduce this concept in Section 4.4 in the form of the Fisher information matrix. The Fisher information matrix depends on the distribution of the observations. As examples, the expressions for the Fisher information matrix are derived for observations that are normally, Poisson, or multinomially distributed or have a distribution that is an exponential family. The inflow of Fisher information—that is, the contribution of each additional observation to the Fisher information matrix— is also addressed.

The inverse of the Fisher information matrix is called Cramér -Rao lower bound matrix. Under general conditions, unbiased but otherwise unspecified estimators cannot have a variance smaller than the corresponding diagonal element of the Cramér-Rao lower bound matrix. Like the Fisher information matrix, the Cramér–Rao lower bound matrix is a key notion in measurement since it specifies a bound to the precision of unbiased parameter measurement. This is the reason why, in Sections 4.5 and 4.6, we present a detailed account of the Cramér–Rao lower bound matrix and its properties. Also, in Section 4.7, we derive expressions for the Fisher information matrix and the Cramér–Rao lower bound matrix for complex parameters. They simplify applications of these concepts in measurement of complex quantities. In Section 4.8, an expression for the Cramér–Rao lower bound matrix for exponential family distributed observations is derived. Generally, the Cramér–Rao lower bound matrix exists only if the corresponding Fisher information matrix is nonsingular. Then, the parameters are called identifiable. Identifiability is discussed in Section 4.9.

The various expressions for the Cramér-Rao lower bound matrix emerging in this chapter show that this bound is typically a function of variables that may, within certain bounds, be freely chosen by the experimenter. An example are the measurement points. This offers the opportunity to select these free variables so that the bound is influenced in a desired way. This technique, called experimental design, is introduced and explained in Section 4.10. An example is the minimization of the Cramér-Rao lower bound on the variance with which a particular parameter may be estimated. Even if such an optimal design itself is not used, it may act as a reference to which the nonoptimal experimental design used or preferred by the experimenter may be compared.

The Cramér-Rao lower bound presents a limit to the precision of unbiased estimators. It does not indicate how to find estimators that are precise, in the sense of having a precision more or less comparable to the Cramér–Rao lower bound. Also, it does not inform about inaccuracy, that is, bias. These questions are addressed in Chapter 5, devoted to precise and accurate estimation.

After a short introduction in Section 5.1, maximum likelihood estimators are introduced in Section 5.2. Under general conditions, these have attractive properties described in Section 5.3. For us, the most important of these is that, typically, the variance of the maximum likelihood estimator attains the Cramér–Rao lower bound if the number of observations used is large enough. Therefore, under this condition, the maximum likelihood estimator may be rightly called most precise. The maximum likelihood estimators of the parameters of the expectation model for normally, Poisson, and multinomially distributed observations are derived and discussed in Sections 5.4–5.6, and those for exponential family distributed observations are covered in Section 5.7. Earlier in this chapter, we presented an example of a multiexponential decay model with and without an additional parameter representing the background or, equivalently, with a background parameter different from or equal to zero.

In Section 5.8, a statistical test is presented enabling the experimenter to conclude from the available observations if there is reason to reject constraints on the parameters such as an equality to zero. This test, the likelihood ratio test, is subsequently specialized to testing if the expectation model used must be rejected. For exponential families of distributions, a simple general expression is derived for the likelihood ratio used in the latter test.

For normally distributed observations, the maximum likelihood estimator is equivalent to the weighted least squares estimator with the elements of the inverse of the covariance matrix of the observations as weights. However, in practice, the least squares estimator is widely applied to observations of any distribution. This is an additional reason why, in Sections 5.9–5.19, extensive attention is paid to it. After an introduction to least squares estimation in Section 5.9, we discuss in Section 5.10 nonlinear least squares estimation. This is least squares estimation applied to expectation models that are nonlinear in one or more of their parameters. Typically, nonlinear least squares estimators are not closed-form and require iterative numerical treatment. Sections 5.11–5.19 deal with various aspects of linear least squares estimation. This is least squares estimation applied to expectation models that are linear in all parameters. An essential difference with nonlinear least squares is that the estimator is now a closed-form expression. In Section 5.12, we first present the general solution for weighted linear least squares estimation with arbitrary weighting matrix. The most important properties of this estimator are that it is unbiased and linear in the observations. Then, the weighting matrix is presented that for any distribution of the observations yields the most precise weighted linear least squares estimator. It is called the best linear unbiased estimator. Chapter 5 is concluded by introducing recursive linear least squares estimators. These update the value of the estimates of the parameters with each additional observation. Two different versions of this estimator are presented in Section 5.18 and Section 5.19, respectively. The first is an ordinary least squares estimator, that is, its weighting matrix is the identity matrix. The second is suitable for tracking time varying parameters since the weighting matrix is chosen so that observations influence the estimates more as they are more recent.

In the final chapter, Chapter 6, we explain principles and use of iterative numerical function optimization methods relevant to the estimators or experimental designs described in this book. The estimators require either the log-likelihood function to be maximized or the least squares criterion to be minimized. The experimental designs require the Cramér–Rao lower bound matrix to be optimized in the sense of a chosen optimality criterion. These optimization problems have in common that their solution is typically not closed-form and has to be computed iteratively.

After a short introduction in Section 6.1, key notions in numerical function optimization are introduced in Section 6.2. These are: the objective function, which is the function to be optimized; the gradient vector, which is the vector of first-order partial derivatives of the objective function; and the Hessian matrix, which is the matrix of second-order derivatives with respect to the independent variables. In the optimization problems in this book, the objective functions are the log-likelihood function or the least squares criterion as a function of the parameters, or the optimality criterion for the experimental design as a function of free experimental variables. Furthermore, the concept ascent (descent) direction of a vector is defined. Finally, the concepts exact observations, reference log-likelihood function, and reference least squares criterion are introduced in Section 6.2. These facilitate software testing.

Section 6.3 is devoted to the steepest descent (ascent) method. This is a general function minimization (maximization) method. It is not specialized to optimizing least squares criteria or log-likelihood functions. The method converges under general conditions, but its rate of convergence may be insufficient. This is improved by the Newton method discussed in Section 6.4. This is also a general function optimization method, but the conditions under which it converges are less general than those for the steepest descent method.

In Section 6.5, the Fisher scoring method is introduced. Its iteration step is an approximation to the Newton step for maximizing log-likelihood functions of parameters of expectation models. Therefore, the Fisher scoring method is a specialized method.

In Section 6.6, an expression is derived for the Newton step for maximizing the log-likelihood function for normally distributed observations. Because of the particular form of the log-likelihood function concerned, this step is also the Newton step for minimizing the nonlinear least squares criterion for observations of any distribution.

From the Newton step for normal observations, a much simpler approximate step may be derived. The method using this step is called Gauss–Newton method and is the subject of Section 6.7. The Newton steps for maximizing the Poisson and the multinomial log-likelihood functions are discussed in Section 6.8 and Section 6.9, respectively. In Section 6.10, an expression is derived for the Newton step for maximizing the log-likelihood function if the distribution of the observations is an exponential family.

From the Newton step for maximizing the log-likelihood function for exponential families of distributions, a much simpler approximate step may be derived that is used by the generalized Gauss–Newton method. This method is the subject of Section 6.11. In Section 6.12, the iteratively reweighted least squares method is described and it is shown that it is identical to the generalized Gauss–Newton method and the Fisher scoring method if the distribution of the observations is a linear exponential family.

Like the Newton method, the Gauss–Newton method solves a system of linear equations for the step in each iteration. The Levenberg–Marquardt method, discussed in Section 6.13, is a version of the Gauss–Newton method that can cope with (near-)singularity of these equations that could occur during the iteration process.

Section 6.14 is a summary of the numerical optimization methods discussed.

Finally, Section 6.15 is devoted to parameter estimation methodology. A number of intermediate steps are recommended in the process starting with choosing a statistical model of the observations and ending with actually estimating the parameters from experimental observations.