MEASUREMENT ERRORS

In applied science and engineering, it is agreed that all observations contain errors. This article discusses how these errors are described today. It also discusses how this description is used to compute the effect of the errors upon the measurement result, and how this effect is reduced or even minimized by rational use of the observations and by experimental design.

For the description of observations, use is made of a mathematical model. This is a mathematical expression intended to describe the observations fully. It will be supposed throughout this article that the mathematical model is parametric, and that the parameters are the quantities to be measured. For example, the parametric model may be a sinusoidal function of time with unknown amplitude, frequency, and phase. Then, these three quantities are the parameters of this model. Yet the model of the observations is incomplete without including errors. If there is reason to assume that the errors in the observations are nonsystematic and additive, they are taken into consideration by adding a term representing them to the expression for the sinusoidal function. Then, the resulting sum is the mathematical model of the observations. It is supposed to fully describe the observations. Nonsystematic errors may loosely be defined as errors that vary if the experiment is repeated under the same conditions and are equal to zero if averaged over many experiments. They are modeled as stochastic variables with an expectation equal to zero. The term expectation is the abstract mathematical term for the mean value. It will be used throughout to avoid confusion with averages, such as time averages, which are measurements. If, in the example, the errors are stochastic variables, so are the observations. Since each observation is equal to the sum of the sinusoidal function and the stochastic error, the expectation of an observation is the value of the function at the time instant concerned. Therefore, the expectation or, equivalently, the function value represents the hypothetical errorless observation.

Since the observations are stochastic variables, they are described by probability density functions. These define for discrete stochastic variables, such as counting results, the probability of occurrence of a particular discrete outcome. For continuous stochastic variables, the probability density function defines the probability of occurrence of an observation within a particular range of values. The probability density function also determines the expectation of the observations. Since this expectation is equal to the function value, the probability density function of the observations depends on the parameters of the function, and thus, the measurement problem has become a statistical parameter estimation problem. This observation has important consequences. It implies that for measurement, use can be made of the extensive theory and methods of statistics. It will be seen that this offers a number of exceptional advantages. A description of these advantages requires some familiarity with a number of notions from statistics. Therefore, these will first be introduced. References 1–4 are useful general texts on statistics.

In statistics, the function of the observations with which a parameter is estimated is called an estimator. Using the same observations, for a particular parameter, different estimators can be defined. Since the observations are stochastic variables, so is the estimator. Therefore, the estimator has a probability density function, an expectation, and a standard deviation. If the expectation of the estimator is equal to the hypothetical true value of the parameter to be estimated, the estimator is called unbiased. Otherwise, it is biased. The deviation of the true value from the expectation is called bias.

Bias is the systematic error. It is, therefore, equivalent to the concept accuracy. There are two essentially different

sources of bias. In the first place, the expectation of the observations may be different from the function model assumed. In the above mentioned sinusoidal example, a trend may be present in addition to the sinusoid, while the model assumed and fitted to the observations consists of the sinusoidal function only. This will, of course, always result in a systematic deviation of the estimated parameters, even in the hypothetical complete absence of nonsystematic errors. The remedy is to include the trend in the model fitted. The inclusion of the trend has the effect that two additional parameters, the slope and the intercept of the trend, have to be estimated. It will be discussed later that there are various reasons to keep in measurement the number of parameters to be estimated as small as possible. Therefore, classical measurement measures to avoid errors such as trends, day-and-night cycles, and background are always preferable to including these contributions in the model. The second source of bias is of a completely different nature. It is produced by and is characteristic of the estimator, itself. It may, therefore, also occur if the assumed model of the observations and that of the expectations are the same. Two estimators of the same parameters from the same observations may have different bias. If the bias vanishes as the number of observations increases, the estimator is called asymptotically unbiased. An effective method to remove bias of this kind is described by Stuart and Ord (4).

The standard deviation of an estimator represents precision. It is the spread of the measurement result if the experiment is repeated under the same conditions. In the example of the estimation of the parameters of the sinusoid, the amplitude, frequency, and phase have to be estimated simultaneously. This estimator is, therefore, vector valued. A vector estimator has a covariance matrix associated with it. The diagonal elements of this matrix are the variances of the estimators of each of the elements of the vector of parameters. The off-diagonal elements represent the covariances of the estimators of different elements. Bias and standard deviation are statistical key properties of estimators for practical measurement purposes. They demonstrate the practical feasibility, clarity, and generality of the model based statistical parameter estimation approach to the treatment of errors in observations. In addition to these desirable properties, model based parameter estimation has a number of advantages which will now be discussed.

It has been mentioned earlier that for the measurement of the parameters of the same model from the same set of observations, use may be made of different estimators. These estimators will generally have different standard deviations, that is, have different precision. The question may, therefore, be asked which estimator is most precise? Or put somewhat differently, what precision is attainable if any estimator may be used? The answer to this question may be given using the concept of Fisher information. The Fisher information with respect to the model parameters is computed from the probability density function of the observations. If the model has one parameter, the quantity computed is called the Fisher information amount. If more than one parameter is measured, such as in the sinusoidal example, it assumes the form of a matrix, the Fisher information matrix. This is a symmetrical matrix of the same order as the number of parameters. The elements of the Fisher information matrix are dependent on the probability density of the observations and on the model of the expectations. They are independent of any method of estimation. With any set of observations, considered as stochastic variables, a Fisher information matrix with respect to the unknown parameters is associated. The inverse of the Fisher information matrix is called the Cramér Rao lower bound. It can be shown that the diagonal elements of the covariance matrix of any unbiased vector estimator, the variances of the elements, cannot be smaller than the corresponding diagonal elements of the Cramér Rao lower bound. Therefore, any unbiased estimator is at best as precise as a hypothetical estimator of the same parameters having the Cramér Rao lower bound as its covariance matrix. Thus, the Cramér Rao lower bound is a standard to which the precision of an unbiased estimator may be compared. This is the reason why the ratio of a diagonal element of the Cramér Rao lower bound to the variance of an estimator is called the efficiency of the estimator. Cramér Rao theory also extends to functions of the estimated parameters. For example, suppose that the height, width, and location of a Gaussian pulse or spectral peak are estimated, but that the quantities to be measured ultimately are the location and the area. Then, the Cramér Rao lower bound for the height, width, and location combined with the expression for the area in terms of the width and height may be used to compute the Cramér Rao lower bound for the location and area. The resulting expressions exactly describe the propagation of the Cramér Rao lower bound for the original parameters to that for location and area. Thus, they show exactly the sensitivity of the Cramér Rao lower bound for the area parameter to the various elements of the Cramér Rao lower bound for the original parameters. This means that an instrument has been found to compute the propagation of errors in the observations to errors in the parameters and, subsequently, to errors in functions of the parameters.

For the engineer and applied scientist, an important question is how to make the influence of errors in the observations upon the measurement result as small as possible. This is equivalent to the question of how to find the method that produces the most precise measurement result from the available observations. For calibration purposes, precision itself may be the ultimate purpose. In other applications, precision is often pursued not for its own sake but to make the conclusions drawn from the measurement result more reliable. The extent to which it is possible to find an efficient estimation method depends on the available a priori knowledge of the probability density function of the observations. As has been discussed earlier, this probability density function is parametric in the hypothetical exact values of the unknown parameters. This dependence of the probability density function of the observations on the parameters may be used to derive a so-called maximum likelihood estimator of the parameters. First, the numerical values of the available observations are substituted for the corresponding independent variables of the probability density function. Next, the true values of the parameters are considered to be variables. The function thus obtained is called the likelihood function of the parameters. Finally, the likelihood function is maximized with respect to the parameters. The values of the parameters at the maximum are the maximum likelihood estimates of the parameters. This procedure shows the first advantage of the maximum likelihood estimator: it is easily found. In addition, the maximum likelihood estimator has a number of favorable statistical properties. The most important of these is that under general condi-

tions, it attains asymptotically the Cramér Rao lower bound. This means that for a large number of observations, it is most precise. The elements of the Cramér Rao lower bound depend on experimental variables. For example, in the estimation of the parameters of the sinusoid and those of the Gaussian pulse, these are the number of observations and their location. If experimental variables such as these may to a certain extent be freely chosen, this freedom may be used to minimize the Cramér Rao lower bound and, thus, the asymptotic variance. This manipulation of the covariance matrix using experimental variables is called experimental design. From a practical point of view, experimental design may be very attractive since it may lead to a more precise measurement result with the same effort or even less. For practical measurement, the invariance property of maximum likelihood estimators is also important: functions of maximum likelihood estimators are maximum likelihood estimators themselves. Generally, maximizing the likelihood function with respect to the parameters is a nonlinear optimization problem which can only be solved using an iterative numerical optimization method. For a long time, this has been a serious impediment to the application of maximum likelihood, but today, excellent optimization methods and software are available which make the method accessible to any user.

If the observations are normally distributed, the maximum likelihood estimator of the parameters can be shown to be the weighted least squares estimator with the inverse of the covariance matrix of the observations as weighting matrix. If the observations are linear in all parameters to be estimated, the least squares estimator is a relatively simple closed form expression linear in the observations. In addition, if the observations are not normally distributed, the weighted least squares method with the inverse covariance matrix as weighting matrix still has the smallest variance among all estimators that are both linear in the observations and unbiased. If the observations are nonlinear in one or more of the parameters to be estimated, the least squares estimator is, as a rule, no longer a closed form and has to be evaluated using an iterative numerical method. However, effective, specialized, and reliable numerical methods and software are available that make the use of nonlinear least squares straightforwardly. As a result, least squares has become a major tool in the handling of observations subject to error in general and not only of normally distributed observations.

EXPECTATIONS OF OBSERVATIONS

The reduction or minimization of the effect of errors in the observations upon the measurement result requires a mathematical model of the observations. In this article, additive nonsystematic errors in the observations will be modeled as stochastic variables with an expectation equal to zero. This implies that the observations are also stochastic variables, and that the expectations of the observations are the hypothetical exact or errorless observations. Thus, these expectations constitute the model underlying the observations. It will be assumed throughout that this model is a parametric function, and that parameters of this model are the quantities to be measured, or that the quantity to be measured can be computed from these parameters.

Example 1. The multiexponential model. Multiexponential observations are observations with expectations

$$y_n(\gamma) = \alpha_1 \exp(-\beta_1 x_n) + \ldots + \alpha_L \exp(-\beta_L x_n) \tag{1}$$

with $n = 1, \ldots, N$ where N is the number of observations, and the $2L \times 1$ vector γ is defined as $(\alpha_1 \ldots \alpha_L \; \beta_1 \ldots \beta_L)^{\mathrm{T}}$, where the amplitudes α_ℓ and the decay constants β_ℓ are the parameters to be measured, and the superscript T denotes transposition. The measurement points x_n, $n = 1, \ldots, N$ are supposed known. If, different from Eq. (1), there is a linear trend in the observations, this deterministic contribution has to included in the model of the expectations of the observations

$$y_n(\eta) = \alpha_1 \exp(-\beta_1 x_n) + \ldots + \alpha_L \exp(-\beta_L x_n) + \lambda x_n + \mu$$

where $\eta = (\alpha_1 \ldots \alpha_L \; \beta_1 \ldots \beta_L \; \lambda \; \mu)^{\mathrm{T}}$. In this expression, λ and μ are the slope and the intercept of the trend, respectively. These parameters have to be estimated along with the parameters α_ℓ and β_ℓ. This means that the number of parameters to be estimated has increased by two. It will be shown below that this is not only disadvantageous from a computational point of view, but it also unfavorably influences the precision with which the α_ℓ and β_ℓ can be measured. Therefore, it is worthwhile to keep the number of parameters as small as possible. As a consequence, changing the experimental conditions to remove the trend is always preferable to including it in the model of the expectations. On the other hand, if trends cannot be avoided, they have to be included since otherwise, the model of the expectations is wrong. Then, values for the amplitudes and decay constants are found systematically deviating from the α_ℓ and β_ℓ, even in the hypothetical case that nonsystematic errors in the observations are absent.

THE DISTRIBUTION OF THE OBSERVATIONS

The mathematical model of the observations is completed by a description of how the observations are distributed about their expectations. This is done in the form of the joint probability density function of the observations. If $w = (w_1 \ldots w_N)^{\mathrm{T}}$ is the vector of the N available observations, their probability density function may be described as $p(w)$. Then, the expectation $\mathrm{E}[w] = (\mathrm{E}[w_1] \ldots \mathrm{E}[w_N])^{\mathrm{T}}$ is defined as

$$\mathrm{E}[w] = \int \ldots \int w p(w) dw \tag{2}$$

where $dw = dw_1 dw_2 \ldots dw_N$, and the integrations are carried out over all possible values of w. Then,

$$\mathrm{E}[w] = y(\theta) \tag{3}$$

where $y(\theta) = [y_1(\theta) \ldots y_N(\theta)]^{\mathrm{T}}$, and $y_n(\theta)$ is the function parametric in the unknown parameters θ defining the errorless observations such as the exponential model described by Eq. (1). For engineering and applied science, two probability density functions are particularly important. These are the normal probability density function and the Poisson probability density function.

Example 2. The normal probability density function. The observations $w_1, \ldots, w_N$ are said to be normally distrib-

uted if their probability density function is described by

$$p(w) = \frac{1}{(2\pi)^{N/2}(\det W)^{1/2}} \exp\left[-\frac{1}{2}(w - \mathrm{E}[w])^{\mathrm{T}} W^{-1}(w - \mathrm{E}[w])\right]$$

where the $N \times N$ matrix W is the covariance matrix of the observations defined by its (i, j)-the element $\mathrm{cov}(w_i, w_j)$ and $\det W$ and W^{-1} are the determinant and the inverse of W, respectively. This probability density function and many others are discussed in Ref. 3. Equation (3) defines the functional dependence of the normal probability density function on the parameters of the function modelling the expectations. For what follows, the logarithm of $p(w)$ as a function of the parameters θ is needed. After substituting $y(\theta)$ for $\mathrm{E}[w]$, it is described by

$$-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det W) - \frac{1}{2}[w - y(\theta)]^{\mathrm{T}} W^{-1}[w - y(\theta)] \quad (4)$$

Notice that both first terms of this expression are independent of the parameter vector θ, while the last term is a quadratic form in the elements $w_n - y_n(\theta)$ of $w - y(\theta)$. Observations in practice are often, but not always, normally distributed. One of the reasons is that if the nonsystematic errors are the sum of a number of nonsystematic errors from independent sources, their distribution tends to normal as described by the central limit theorem discussed in Ref. 2.

Example 3. The Poisson probability density function. This probability density function concerns counting statistics. It is described in Ref. 3. Examples of Poisson distributed stochastic variables are radioactive particle counts and pixel values in electron microscopes. The number of counts is Poisson distributed if the probability that it is equal to w_n is given by

$$p(w_n) = \exp(-\lambda_n)\frac{\lambda_n^{w_n}}{w_n!} \quad (5)$$

Simple calculations show that $\mathrm{E}[w_n] = \lambda_n$, and that the standard deviation of w_n is equal to $\sqrt{\lambda_n}$. If $w_1, \ldots, w_N$ are independent, as is often assumed in applications, their joint probability density function is equal to the product the probabilities described by Eq. (5)

$$p(w) = \prod_n p(w_n) \quad (6)$$

with $n = 1, \ldots, N$. Since $\mathrm{E}[w] = \lambda$ with $\lambda = (\lambda_1 \ldots \lambda_N)^{\mathrm{T}}$ and $\lambda_n = y_n(\theta)$, the logarithm of the probability density function defined by Eq. (6) may be written

$$\sum_n -y_n(\theta) + w_n \ln[y_n(\theta)] - \ln(w_n!) \quad (7)$$

Notice that the last term in this expression is independent of the parameter vector θ.

From Example 2 and Example 3, the general approach to establishing the dependence of the probability density function of the observations on the parameters, that is, the quantities to be measured, is now clear. First, the expectation of the observations w_n is computed. Then, the result is substituted for the relevant quantities in the probability density function. The probability density function thus obtained is parametric in the parameters of the expectations, that is, of the hypothetical errorless observations. This is the form of the probability density function that will be used hereafter for two purposes. First, it will be used for the computation of the highest attainable precision with which the parameters can be measured from the available observations. It will also be used to find the most precise method to estimate the parameters from the observations.

ATTAINABLE MEASUREMENT PRECISION IN THE PRESENCE OF MEASUREMENT ERRORS

Suppose that a number of N observations $w_1, \ldots, w_N$ is available and that the expectations of the observations are described by the multiexponential model defined by Eq. (1). If this model is fitted to the observations with respect to its parameters, the amplitudes, and the decay constants, one could choose the sum of the squares of the deviations of the model from the observations as a criterion of goodness of fit. Then, this criterion could be minimized with respect to the parameters, and the parameter values for which the criterion would be minimum would be the solution. This is the well-known ordinary least squares solution. Alternatively, one could have chosen the values of the parameters for which the sum of the absolute values of the deviations would be minimum. This is the least absolute values or least moduli solution. Then, if the experiment could be repeated sufficiently often, the experimenter could compare the results of both methods and could decide which of both would be most precise. Seeing that the one method is more precise than the other, the experimenter might wonder what the highest attainable precision from these observations is with any method. It has been found that under general conditions, this question may be answered using the concept Fisher information. For a discussion of Fisher information, see Ref. 4. For the computation of the Fisher information, the probability density function of the observations $p(w;\theta)$ is used. This is done as follows. First, the logarithm of $p(w;\theta)$ is taken. For the normal probability density function and for the Poisson probability density function, the result of this operation is described by Eqs. (4) and (7), respectively. Next, the gradient vector of $\ln p(w;\theta)$ with respect to the elements of θ is calculated. It is defined as

$$\frac{\partial \ln p(w;\theta)}{\partial \theta}$$

If θ is a $K \times 1$ vector, so is the gradient vector. Its k-th element is $\partial \ln p(w;\theta)/\partial\theta_k$. Next, the $K \times K$ matrix

$$\frac{\partial \ln p}{\partial \theta}\frac{\partial \ln p}{\partial \theta^{\mathrm{T}}} \quad (8)$$

is computed where, for simplicity, the arguments of $p(w;\theta)$ have been left out, and $\partial \ln p/\partial\theta^{\mathrm{T}}$ is the transpose of $\partial \ln p/\partial\theta$. The $K \times K$ Fisher information matrix is defined as the expectation of Eq. (8)

$$M = \mathrm{E}\left[\frac{\partial \ln p}{\partial \theta}\frac{\partial \ln p}{\partial \theta^T}\right] = \int \cdots \int \frac{\partial \ln p}{\partial \theta}\frac{\partial \ln p}{\partial \theta^T} p\, dw$$

It is not difficult to show that M may alternatively be written

$$M = -\mathrm{E}\left[\frac{\partial^2 \ln p}{\partial\theta\,\partial\theta^{\mathrm{T}}}\right] \tag{9}$$

In this expression, $\partial^2 \ln p/\partial\theta\,\partial\theta^{\mathrm{T}}$ is the Hessian matrix of $\ln p$ defined by its (q, r)-th element $\partial^2 \ln p/\partial\theta_q\,\partial\theta_r$.

Example 4. The Fisher information matrix for normally distributed observations. If the observations are normally distributed, the logarithm of the probability density function as a function of the parameters is described by Eq. (4). Then, elementary computations making use of the fact that $\mathrm{E}[w_n - y_n(\theta)] = 0$ yield

$$M = \frac{\partial y^{\mathrm{T}}}{\partial\theta} W^{-1} \frac{\partial y}{\partial\theta^{\mathrm{T}}} \tag{10}$$

In this expression, the $N \times K$ matrix $\partial y/\partial\theta^{\mathrm{T}}$ is the Jacobian matrix of $y(\theta)$ with respect to θ. Its (n, k)-th element is equal to $\partial y_n(\theta)/\partial\theta_k$. Therefore, for the multiexponential model, the elements of the Jacobian matrix are of the form $\exp(-\beta_\ell x_n)$ or $-\alpha_\ell\, x_n \exp(-\beta_\ell x_n)$ with $\ell = 1, \ldots, L$.

Example 5. The Fisher information matrix for independent Poisson distributed observations. If the observations are independent and Poisson distributed, the logarithm of the probability density function of the observations as a function of the parameters is described by Eq. (7). This expression may be used to show that here, the information matrix is also described by Eq. (10), but with $W = \mathrm{diag}(y_1 \ldots y_N)$, where $y_n = y_n(\theta)$.

The importance of the Fisher information matrix is that from it the Cramér Rao lower bound may be computed. This is a lower bound on the variance of all unbiased estimators of parameters or of functions of parameters. An estimator t is said to be unbiased for the parameter θ if its bias, defined as

$$\mathrm{E}[t] - \theta$$

is equal to the null vector. Otherwise, it is biased. In measurement terminology: if the model of the expectations of the observations is correctly specified and the estimator used is unbiased for the parameters, the measurement result has no systematic error.

Next, suppose that $t(w)$ is any unbiased estimator of the vector of parameters θ from the observations w. Then the Cramér Rao inequality states that

$$\mathrm{cov}[t(w), t(w)] \geq M^{-1} \tag{11}$$

In this expression, $\mathrm{cov}[t(w), t(w)]$ is the covariance matrix of the estimator $t(w)$. That is, the (p,q)-th element of this matrix is defined as the covariance of the p-th element $t_p(w)$ and the q-th element $t_q(w)$. Therefore, the diagonal elements are the variances of $t_1(w), \ldots, t_K(w)$, respectively. Ineq. (11) expresses that the difference of the matrix $\mathrm{cov}[t(w), t(w)]$ and the matrix M^{-1} is positive semidefinite. A property of positive semidefinite matrices is that their diagonal elements cannot be negative. Therefore, the diagonal elements of $\mathrm{cov}(t(w), t(w))$, that is, the variances of the elements of the estimator $t(w)$, must be larger than or be equal to the corresponding diagonal elements of M^{-1}. Consequently, the latter diagonal elements are a lower bound on the variances of the elements of the estimator $t(w)$. The matrix M^{-1} is called the Cramér Rao lower bound. For normally distributed observations and for Poisson distributed observations, the Cramér Rao lower bound may be computed by inverting the Fisher information matrix defined by Eq. (10) with appropriate matrix W, respectively. Notice that the main ingredients are simply the derivatives of the model $y_n(\theta)$ with respect to the parameters in each measurement point. These are quantities that are usually easy to compute.

The Cramér Rao lower bound would be of theoretical value only if there would not exist estimators attaining it. Later in this article, estimators will be introduced that do so, at least asymptotically. Therefore, the Cramér Rao lower bound may be used as a standard to which the precision of any estimator may be compared. Notice that the Cramér Rao lower bound is not related to a particular estimation method. It depends on the statistical properties of the observations, the measurement points, and in most cases, the hypothetical true values of the parameters. This dependence on the true values looks, at first sight, as a serious impediment to the practical use of the bound. However, the expressions for the bound provide the means to compute numerical values for it using nominal values of the parameters. This provides the experimenter with quantitative insight in what precision may be achieved from the available observations, an insight that without the bound would be absent. Thus, using the bound, the experimenter gets a detailed insight in the sensitivity of the precision to the values of the parameters. The experimenter also gets impression if the experimental design, that is, the values and the number of the measurement points x_n, is adequate for the purposes concerned. This means an impression if the precision is sufficient to make conclusions possible. If not, there is no other choice than to change the experimental design. If this is not possible, it is to be concluded that the observations are not suitable for the purposes of the measurement procedure.

In many applications, some of the quantities to be measured are functions of the parameters and not the individual parameters. A simple example is the following.

Example 6. Measurement of peak area and location. Suppose that a number of error corrupted observations $w_1, \ldots, w_N$ has been made on a spectral peak described by

$$\alpha \exp\left[-\frac{1}{2}\left(\frac{x-\beta}{\gamma}\right)^2\right] \tag{12}$$

where the parameters α, β, and γ are the peak height, location, and half-width, respectively. Suppose that only the peak location and the peak area are of interest. Then these are described by β and $(2\pi)^{1/2}\alpha\gamma$, respectively.

Fortunately, the Cramér Rao lower bound of functions of the parameters follows relatively easily from the Cramér Rao lower bound for the parameters. Let $r = [r_1(w) \ldots r_L(w)]^{\mathrm{T}}$ be an unbiased estimator of the vector function $\rho(\theta) = [\rho_1(\theta) \ldots \rho_L(\theta)]^{\mathrm{T}}$, that is, $\mathrm{E}[r] = \rho(\theta)$. Furthermore, let M be the informa-

tion matrix for θ. Then it can be shown that

$$\mathrm{cov}[r(w), r(w)] \geq \frac{\partial \rho}{\partial \theta^T} M^{-1} \frac{\partial \rho^{\mathrm{T}}}{\partial \theta}$$

where $\partial\rho/\partial\theta^{\mathrm{T}}$ is the $L \times K$ Jacobian matrix with (p, q)-th element $\partial\rho_p/\partial\theta_q$. Therefore, the Cramér Rao lower bound for unbiased estimation of ρ is described by

$$\frac{\partial \rho}{\partial \theta^{\mathrm{T}}} M^{-1} \frac{\partial \rho^{\mathrm{T}}}{\partial \theta} \tag{13}$$

with M^{-1} the Cramér Rao lower bound for α, β, and γ.

Example 7. The Cramér Rao lower bound for peak area and location. The vector $\rho(\theta)$ for Example 6 is described by $\rho(\theta) = [\beta\ (2\pi)^{1/2}\alpha\gamma]^{\mathrm{T}}$. Then, the Jacobian matrix of $\rho(\theta)$ with respect to $(\alpha\ \beta\ \gamma)^{\mathrm{T}}$ is defined as

$$\frac{\partial \rho}{\partial \theta^{\mathrm{T}}} = \begin{pmatrix} 0 & 1 & 0 \\ (2\pi)^{1/2}\gamma & 0 & (2\pi)^{1/2}\alpha \end{pmatrix}$$

where $\rho = \rho(\theta)$. The Cramér Rao lower bound for unbiased estimation of ρ is then computed from Eq. (13).

The premultiplication and postmultiplication of M^{-1} in Eq. (13) describe what is conventionally called error propagation. To see how this works, suppose that $\rho = [\rho_1(\theta)\ \rho_2(\theta)]^{\mathrm{T}}$, $\theta = (\theta_1\ \theta_2)^{\mathrm{T}}$, and let the Cramér Rao lower bound for θ be

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{pmatrix}$$

Then, the diagonal elements of the Cramér Rao lower bound for ρ_1 and ρ_2 are equal to

$$\left(\frac{\partial \rho_i}{\partial \theta_1}\right)^2 c_{11} + 2\left(\frac{\partial \rho_i}{\partial \theta_1}\right)\left(\frac{\partial \rho_i}{\partial \theta_2}\right) c_{12} + \left(\frac{\partial \rho_i}{\partial \theta_2}\right)^2 c_{22}$$

with $i = 1, 2$, respectively. This expression shows how the variances c_{11} and c_{22} and the covariance c_{12} of a hypothetical estimator that attains the Cramér Rao lower bound for θ propagate to the variances of a hypothetical estimator of ρ_1 and ρ_2 that also attains the Cramér Rao lower bound. Similar error propagation schemes are proposed in the literature for covariance matrices of functions of estimators in general, for example in reference 5. These schemes are approximations using the linear Taylor polynomial instead of the nonlinear functions. Equation (13), on the other hand, is exact.

Next, suppose that M is the information matrix for the estimation of $\theta = (\theta_1 \ldots \theta_K)^{\mathrm{T}}$, and that an additional parameter θ_{K+1} is to be estimated. For example, θ_{K+1} may be a constant term added to the spectroscopic line model described by Eq. (12) to model a constant background contribution. Then, M has to be augmented with one row and one column corresponding to θ_{K+1}. If the augmented information matrix is inverted, all first K diagonal elements can be shown to be larger than or equal to the corresponding diagonal elements of M^{-1}. Equality occurs only if the nondiagonal elements of the $(K+1)$-th row and $(K+1)$-th column of the augmented information matrix happen to be equal to zero. Generally, it is not difficult to show that, typically, the first K diagonal elements of M^{-1} are monotonously increasing with the number of parameters in excess of K.

PRECISELY MEASURING FROM ERROR CORRUPTED OBSERVATIONS

The a priori knowledge of the experimenter about the observations and the extent to which this a priori knowledge is used may considerably influence the precision and accuracy of the measurement result. This concerns both systematic and nonsystematic errors in the observations. Systematic errors in the observations are deviations of the assumed parametric model of the expectations from the true model of the expectations. Even in the absence of nonsystematic errors, discrepancy between both models produces systematic errors, that is, inaccuracy in the measurement result. Since no model fitted will be perfect, there will always be a certain amount of systematic error. Nonsystematic errors are described by their distribution about the expectations of the observations. This distribution is not always known, but if it is, this knowledge may contribute substantially to the reducing of the nonsystematic error in the measurement result, that is, in the parameters estimates.

Suppose that observations $w_1, \ldots w_N$, are available and that their probability density function $f(\omega_1, \ldots, \omega_N;\theta)$ is known where θ is the vector of unknown parameters and $\omega_1, \ldots, \omega_N$ are the independent variables corresponding to the observations $w_1, \ldots, w_N$, respectively. Assume that $w_1, \ldots, w_N$ are substituted for $\omega_1, \ldots, \omega_N$ in $f(\omega_1, \ldots, \omega_N;\theta)$, respectively, and that the fixed true parameters θ are replaced by the vector of corresponding variables t. Then, the resulting function $f(w_1, \ldots, w_N;t)$ of t is called the likelihood function of the parameters t, given the observations $w_1, \ldots, w_N$. The maximum likelihood estimate of the parameters θ is defined as the value $\tilde{t}$ of t that maximizes the likelihood function.

The maximum likelihood estimator has a number of very favorable properties. In the first place, its definition shows that it is simple to find from the known probability density function of the observations. Furthermore, it can be shown to converge under general conditions in a statistically well-defined way to the true values of the parameters as the number of observations increases. Moreover, under general conditions, the covariance matrix of the maximum likelihood estimator approaches asymptotically the Cramér Rao lower bound. Then, the maximum likelihood estimator is asymptotically most precise. Also, a function of a maximum likelihood estimator is the maximum likelihood estimator of the function. This is called the invariance property of maximum likelihood.

Two of these properties are asymptotic; they apply to an infinite number of observations. If they also apply to a finite or even small number of observations can often only be assessed by estimating from artificial, simulated observations. These simulations may reveal that maximum likelihood estimation applied to small numbers of observations may lead to bias, that is, systematic error in the measurement result. This kind of bias, or the major part of it, is usually inversely proportional to the number of observations and may be removed as follows. Let $\tilde{t}_N$ be the biased maximum likelihood estimate of θ obtained from $w_1, \ldots, w_N$, and let $\bar{t}_{N-1}$ be the

average of the N different maximum likelihood estimates computed from the N different sets of $N-1$ observations obtained by omitting one observation from the set $w_1, \ldots, w_N$. Then, it may be shown that

$$N\tilde{t}_N - (N-1)\bar{t}_{N-1}$$

is an estimator of θ which may only be biased to order $1/N^2$. This is the so-called Quenouille correction, today called jackknife. A favorable property of this correction is that it hardly affects the variance of the estimator.

Example 8. Maximum likelihood estimation of peak height, width and location from Poisson distributed observations. Suppose that observations $w_1, \ldots, w_N$ are available made on the spectral peak of Example 6, and that these observations are independent and Poisson distributed. Then it follows from Example 3 that the likelihood function of the parameters is described by

$$\sum_n -y_n(t) + w_n \ln[y_n(t)] - \ln(w_n!) \tag{14}$$

with

$$y_n(t) = a \exp\left[-\frac{1}{2}\left(\frac{x_n - b}{c}\right)^2\right]$$

where $t = (a\ b\ c)^{\mathrm{T}}$. To obtain the maximum likelihood estimate of α, β, and γ, Eq. (14) must be maximized with respect to t. This is a nonlinear optimization problem which has to be solved numerically. If the peak area is computed from the maximum likelihood estimates $\tilde{a}$ and $\tilde{c}$ as $(2\pi)^{1/2}\tilde{a}\tilde{c}$ this is, by the invariance property, a maximum likelihood estimate as well.

Example 9. Maximum likelihood estimation from observations disturbed by normally distributed errors. If the errors and, therefore, the observations are normally distributed, Eq. (4) shows that the likelihood function of the parameters is described by

$$-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det W) - \frac{1}{2}[w - y(t)]^T W^{-1}[w - y(t)] \tag{15}$$

Since both first terms of this expression do not depend on the vector of parameters t, maximizing Eq. (15) is equivalent to minimizing

$$[w - y(t)]^{\mathrm{T}} W^{-1}[w - y(t)]$$

with respect to t. This shows that with normally distributed observations, maximum likelihood estimation is equivalent to a weighted least squares measurement with W^{-1} as weighting matrix.

Least squares estimation is also often used if the distribution of the observations is not known or is known to be not normal. Then, the general expression for the least squares criterion is

$$[w - y(t)]^{\mathrm{T}} \Omega [w - y(t)] \tag{16}$$

where Ω is a positive definite weighting matrix to be chosen by the experimenter.

Linear Least Squares

First, as an important special case, models linear in the unknown parameters θ are considered. Then

$$\mathrm{E}[w] = y(\theta) = X\theta$$

with X a known $N \times K$ matrix, that is,

$$y_n(\theta) = x_{n1}\theta_1 + \ldots + x_{nK}\theta_K$$

Notice that

$$x_n = (x_{n1} \ldots x_{nK})^{\mathrm{T}}$$

is the vector independent variable corresponding to the n-th observation w_n.

Example 10. Straight line fitting. If the observations w_n are made on a straight line $y = \alpha x + \beta$ at the points $x_{11}, \ldots, x_{N1}$, then X and θ are described by

$$\begin{bmatrix} x_{11} & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ x_{N1} & 1 \end{bmatrix} \quad \text{and} \quad \theta = (\alpha\ \beta)^{\mathrm{T}}$$

respectively.

The least squares solution $\hat{t}_\Omega$ for θ is

$$\hat{t}_\Omega = (X^{\mathrm{T}}\Omega X)^{-1} X^{\mathrm{T}}\Omega w \tag{17}$$

It is observed that this solution is a linear combination of the observations. As a result, the propagation of the errors in the observations to the measurement result is perfectly clear. Furthermore, since $\mathrm{E}[w] = X\theta$, $\mathrm{E}[\hat{t}_\Omega] = \theta$ and, hence, $\hat{t}_\Omega$ is an unbiased estimator of θ. Notice that $\hat{t}_\Omega$ has these properties for any distribution of the observations w. It is easily shown that the covariance matrix $\mathrm{cov}(\hat{t}_\Omega, \hat{t}_\Omega)$ is equal to

$$(X^{\mathrm{T}}\Omega X)^{-1} X^{\mathrm{T}}\Omega W \Omega X (X^{\mathrm{T}}\Omega X)^{-1} \tag{18}$$

The conclusion from this expression is that this covariance matrix and, therefore, the variances of $\hat{t}_\Omega$ depend on the choice of Ω. The question is then which Ω minimizes the covariance described by Eq. (18). The answer has been found to be $\Omega = W^{-1}$. For this choice,

$$\mathrm{cov}(\hat{t}_\Omega, \hat{t}_\Omega) \geq \mathrm{cov}(\hat{t}_{W^{-1}}, \hat{t}_{W^{-1}})$$

As a consequence, the variances of the elements of $\hat{t}_\Omega$ for any choice of Ω are never smaller than those of the corresponding elements of $\hat{t}_{W^{-1}}$. The estimator $\hat{t}_{W^{-1}}$ is called the best linear unbiased estimator. Equation (17) shows that it is described by

$$\hat{t}_{W^{-1}} = (X^{\mathrm{T}}W^{-1}X)^{-1}X^{\mathrm{T}}W^{-1}w$$

Among all estimators that are both linear in the observations and unbiased, it is called best since it has smallest variance.

Notice that only the expectation and the covariance matrix of the observations are specified, not their probability density function. Also notice that $\hat{t}_{W^{-1}}$ is optimal within the class of estimators that are both linear in the observations and unbiased. Therefore, there may be better, that is, more precise estimators among those that are not linear in the observations or are biased.

The covariance matrix of $\hat{t}_{W^{-1}}$ is equal to

$$(X^{\mathrm{T}}W^{-1}X)^{-1} \tag{19}$$

If for normally distributed observations, the maximum likelihood estimator is computed, it is found to be identical to the best linear unbiased estimator $\hat{t}_{W^{-1}}$ and, consequently, to have a covariance matrix equal to the one given by Eq. (19). If next, the Cramér Rao lower bound is computed for the same observations, it is found to coincide with Eq. (19). The conclusion is that for normally distributed observations, the best linear unbiased estimator is identical with the maximum likelihood estimator and attains the Cramér Rao lower bound for any number of observations.

In measurement practice, the weighting matrix Ω of $\hat{t}_{\Omega}$ is often taken as the identity matrix. The reason may be that the covariance matrix W is unknown. Another reason may be the amount and the complexity of numerical computation involved since with $\Omega = I$, the estimator simplifies to the ordinary least squares estimator

$$\hat{t}_I = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}w \tag{20}$$

which is clearly easier to compute than $\hat{t}_{\Omega}$. The corresponding ordinary least squares criterion is described by

$$(w - Xt)^{\mathrm{T}}(w - Xt)$$

which is simply the sum of the squares of the deviations

$$w_n - x_{n1}t_1 - x_{n2}t_2 + \ldots - x_{nK}t_K$$

Notice that $\hat{t}_I$ is only the best linear unbiased estimator if the covariance matrix W is equal to $\sigma^2 I$, that is, if the observations are uncorrelated and have equal variance σ^2. The estimator $\hat{t}_I$ is the maximum likelihood estimator and achieves the Cramér Rao lower bound if, in addition, the observations are normally distributed. Therefore, if these conditions are not met, the use of $\hat{t}_I$ may mean an exchange of precision for simplicity.

Finally, it is emphasized that Eq. (20) is a formal description of the ordinary linear least squares estimator. It is not a recipe for its numerical evaluation. Special numerical methods have been designed taking care of the fact that the set of linear equations described by Eq. (20) may be ill-conditioned. References 6 and 7 provide the details.

Nonlinear Least Squares

Nonlinear least squares is the most frequently used method for estimation of the parameters of nonlinear models. The criterion used is described by

$$[w - y(t)]^{\mathrm{T}}[w - y(t)] = \sum_n [w_n - y(x_n; t)]^2 \tag{21}$$

which is Eq. (16) with weighting matrix $\Omega = I$. Notice that generally, the solution $\hat{t}_I$ for t minimizing the least squares criterion defined by Eq. (21) is only the maximum likelihood estimator if the observations are independent and identically normally distributed. This means normally distributed with covariance matrix $\sigma^2 I$. For other distributions, $\hat{t}_I$ is generally not the maximum likelihood estimate since it does not maximize the pertinent likelihood function. As compared with linear least squares, the amount of theory concerning nonlinear least squares is limited. However, if the observations are independent and identically distributed, then under general conditions, the least squares estimator $\hat{t}_I$ is asymptotically normally distributed with covariance matrix

$$\sigma^2 \left(\frac{\partial y^T}{\partial \theta} \frac{\partial y}{\partial \theta^T} \right)^{-1} \tag{22}$$

where $y = y(\theta)$. This result is due to Jennrich (8). Notice that the computation of this covariance matrix requires the parameters to be known. In practice, this is not the case, and nominal or estimated values are substituted for the exact ones. Also notice that for independent and identically normally distributed observations, Eq. (22) is equal to the Cramér Rao lower bound. The general form of the elements of the matrix $(\partial y^{\mathrm{T}}/\partial\theta)(\partial y/\partial\theta^{\mathrm{T}})$ is

$$\sum_n \frac{\partial y(x_n; \theta)}{\partial \theta_p} \frac{\partial y(x_n; \theta)}{\partial \theta_q}$$

This expression shows the dependence of the elements of this matrix upon the values of the independent variable x. Therefore, if the experimenter has some freedom in the choice of the measurement points, it may be used to manipulate the covariance matrix described by Eq. (22) in a desired way. This usually concerns the diagonal elements, that is, the variances and is an example of experimental design: the manipulation of the variances by selecting free experimental variables.

The gradient of the nonlinear least squares criterion with respect to the parameter vector t is equal to

$$-2 \sum_n [w_n - y_n(t)] \frac{\partial y_n(t)}{\partial t} \tag{23}$$

A necessary condition for a point to be a minimum is that the gradient is equal to the null vector. If Eq. (23) is equated to the null vector, this produces a set of K nonlinear equations in K variables. This set must be solved by an iterative numerical method since, typically, it cannot be solved in closed form. For this problem, specialized numerical methods have been developed. Most frequently used are the Gauss–Newton method and the Levenberg–Marquardt method. These are described in references 8 and 6, respectively. Software for their practical implementation is found in references 6 and 7.

Many nonlinear models in engineering practice are linear in some of their parameters. How this special property may be exploited in nonlinear least squares estimation is illustrated in the following example.

Example 11. Least squares estimation of the parameters of a multiexponential model. Suppose that in a least

squares estimation problem, the model fitted is described by

$$y_n(t) = a_1 \exp(-b_1 x_n) + \ldots + a_L \exp(-b_L x_n)$$

where $t = (a^{\mathrm{T}}\ b^{\mathrm{T}})^{\mathrm{T}}$ with linear parameters a = $(a_1\ .\ .\ .\ a_L)^{\mathrm{T}}$ and nonlinear parameters $b = (b_1\ .\ .\ .\ b_L)^{\mathrm{T}}$. Then Eq. (23) shows that at the minimum of the least squares criterion, the derivatives with respect to the linear parameters a must satisfy

$$\sum_n [w_n - a_1 \exp(-b_1 x_n) + \ldots + a_L \exp(-b_L x_n)] \exp(-b_\ell x_n) = 0$$

with $\ell = 1, .\ .\ ., L$. This may be considered a set of L linear equations in L unknowns a_ℓ. The solution for these unknowns is a function of the unknown nonlinear parameters b_ℓ and is denoted as $a_\ell(b)$. Substitution of the $a_\ell(b)$ for the a_ℓ in the least squares criterion yields

$$\sum_n [w_n - a_1(b) \exp(-b_1 x_n) + \ldots + a_L(b) \exp(-b_L x_n)]^2$$

Thus, the least squares criterion has become a function of the nonlinear parameters b only. Minimization of it with respect to b yields the solution $\hat{b}$ for β and the solution $\hat{a} = a(\hat{b})$ for α.

Nonlinear least squares problems of the kind described in Example 11 are called separable nonlinear least squares problems since the linear and the nonlinear parameters are estimated separately. Notice that in Example 11, the number of parameters involved in the iterative numerical minimization is reduced by a factor of two. This also means that the number of initial values for the procedure is reduced correspondingly.

HANDLING MEASUREMENT ERRORS IN NONSTANDARD PROBLEMS

Complex Parameter Estimation

Many practical measurement problems concern complex valued parameters or mixtures of real and complex valued parameters. In particular, these problems are found in measurement in the frequency domain. Such complex parameter estimation problems can always be transformed into real parameter estimation problems by splitting a complex parameter into its real and imaginary part and estimating these real quantities separately. This, however, leads to unnecessarily complicated expressions for the estimator and, as a result, to complicated numerical procedures. This is avoided by leaving quantities complex if they are complex by nature.

The most important tool in the formulation of complex parameter measurement from error corrupted observations is the following. Suppose that in a measurement problem there are $K + 2L$ parameters

$$\theta = (\eta_1 \ldots \eta_K\ \alpha_1\ \beta_1 \ldots \alpha_L\ \beta_L)^{\mathrm{T}}$$

of which the η_k are intrinsically real, and the α_ℓ and β_ℓ are the real and imaginary parts of the complex parameters $\gamma_\ell = \alpha_\ell + \mathrm{j}\beta_\ell$ with $j = \sqrt{-1}$. Then γ_ℓ and its complex conjugate γ_ℓ^* on the one hand and α_ℓ and β_ℓ on the other are connected by the linear transformation

$$\begin{bmatrix} \gamma_\ell \\ \gamma_\ell^* \end{bmatrix} = J \begin{bmatrix} \alpha_\ell \\ \beta_\ell \end{bmatrix}$$

where

$$J = \begin{bmatrix} 1 & j \\ 1 & -j \end{bmatrix}$$

Therefore, the mixed real complex parameter vector

$$\zeta = (\eta_1 \ldots \eta_K\ \gamma_1\ \gamma_1^*\ \gamma_L\ \gamma_L^*)^{\mathrm{T}}$$

and θ are connected by

$$\zeta = B_{K+2L}\theta \tag{24}$$

where B_{K+2L} is the $(K + 2L) \times (K + 2L)$ block diagonal matrix

$$B_{K+2L} = \mathrm{diag}(I_K \quad A_{2L})$$

with I_K the identity matrix of order K and A_{2L} the $2L \times 2L$ block diagonal matrix

$$A_{2L} = \mathrm{diag}(J \ldots J)$$

The theory, methods, and techniques presented up to now concerned the estimation of real parameters from error corrupted observations. Using the linear transformation described by Eq. (24), transformation of the pertinent expressions into those for estimating a mixed real complex parameter vector is relatively easy. All that is required is observing the mathematical rules governing linear transformation of coordinates in general. For what follows, it is important to notice that Eq. (24) implies that α_ℓ and β_ℓ are transformed into both γ_ℓ and γ_ℓ^*. Also, the definition of the covariance matrix of a vector of complex stochastic variables is needed. Let z be a vector of stochastic variables. Then, the covariance matrix of z is defined as

$$\mathrm{E}[(z - E[z])(z - E[z])^{\mathrm{H}}]$$

where the superscript H denotes complex conjugate transposition. The Fisher information matrix defined by Eq. (9) after the transformation of parameters described by Eq. (24) is given by

$$M = -\mathrm{E}\left[\frac{\partial^2 \ln f}{\partial \zeta^* \partial \zeta^T}\right]$$

and the corresponding Cramér Rao lower bound on the variance of unbiased estimators of ζ is equal to M^{-1}. Again using Eq. (24), the Cramér Rao lower bound for a vector of real and complex functions $\phi(\zeta)$ of the mixed real complex parameter vector is found to be

$$\frac{\partial \phi}{\partial \zeta^T} M^{-1} \frac{\partial \phi^{\mathrm{H}}}{\partial \zeta^*}$$

where for brevity, the argument of $\phi(\zeta)$ has been omitted. A further example is the weighted least squares estimator de-

fined by Eq. (17). After transformation of θ into ζ and $\hat{t}_\Omega$ into $\hat{z}_\Omega$, respectively, it becomes

$$\hat{z}_\Omega = (R^H \Omega R)^{-1} R^H \Omega w$$

where the complex $N \times (K + 2L)$ matrix R is equal to XB_{K+2L}^{-1}. If Ω is equal to W^{-1}, this is the best linear unbiased estimator. Finally, suppose that the real complex $(P + 2Q) \times 1$ vector of observations u is composed of the elements of the real $(P + 2Q) \times 1$ vector of observations w as follows

$$u = B_{P+2Q} w$$

with $P + 2Q = N$. Then, u is a vector of real and complex observations described by

$$(w_1 \ldots w_P \, w_{P+1} + jw_{P+2} \, w_{P+1} - jw_{P+2} \ldots w_{P+2Q-1} + jw_{P+2Q} \; w_{P+2Q-1} - jw_{P+2Q})^T$$

and

$$\hat{z}_\Psi = (S^H \Psi S)^{-1} S^H \Psi u$$

where S and Ψ are equal to $B_{P+2Q}R$ and $B_{P+2Q}^{-H} \, \Omega \, B_{P+2Q}^{-1}$, respectively. The covariance matrix of the mixed real complex observations is defined as $E[(u - E[u])(u - E[u])^H]$ and is, therefore, equal to $B_{P+2Q} W B_{P+2Q}^H$. Hence, the estimator $\hat{z}_\Psi$ with $\Psi = (B_{P+2Q} W B_{P+2Q}^H)^{-1}$ is the best linear unbiased estimator.

The iterative numerical optimization of likelihood functions and nonlinear least squares criteria of mixed real complex parameters may be carried out directly with respect to the vector of mixed real complex parameters. This is discussed in reference 9. In particular, use may be made of the complex gradient. Specifically, the complex gradient of the logarithm of the likelihood function $\ln f$ with respect to the complex parameter vector z is defined as $\partial \ln f / \partial z$. An important property of this complex gradient is that the real gradient $\partial \ln f / \partial t$ is equal to the null vector if and only if the complex gradient is equal to the null vector. Therefore, the complex gradient may be used to find maxima of the likelihood function and minima of the nonlinear least squares criterion in the same way as the real gradient.

Nonstandard Fourier Analysis

Estimation of Fourier coefficients from error disturbed observations made on periodic functions is an important problem in dynamic system identification in general and in specialized applications as crystal structure reconstruction. Suppose that the problem is to estimate the Fourier coefficients γ_k, $k = 0, \pm 1, \ldots, \pm K$ and, possibly, the period δ of the real periodic function

$$y_n(\zeta) = \sum_k \gamma_k \exp(-j2\pi k x_n / \delta) \tag{25}$$

from error corrupted observations $w = (w_1 \ldots w_N)^T$ where the measurement points x_n are known, and the vector of unknown parameters ζ is either equal to $\gamma = (\gamma_0 \; \gamma_1 \; \gamma_{-1} \ldots \gamma_K \; \gamma_{-K})^T$ or to $(\gamma^T \; \delta)^T$, where γ_0 and δ are real, while the remaining γ_k are complex and satisfy $\gamma_{-k} = \gamma_k^*$ since the $y_n(\zeta)$ are real. It will not be supposed that the measurement points are equidistant, nor if they are, that the period is a known integer multiple of the sampling interval, and an integer number of periods is observed. The purpose of this section is to formulate the estimation of the parameters ζ as a complex statistical parameter estimation problem and to describe the special conditions under which this problem simplifies to the standard Discrete Fourier Transform, the DFT.

Example 12. Estimation of Fourier coefficients from Poisson distributed observations. Suppose that observations $w_n \geq 0$, $n = 1, \ldots, N$ are available with expectations described by Eq. (25) and that these observations have a Poisson distribution. Then, by Eq. (14), the likelihood function of the parameters is

$$\sum_n -y_n(z) + w_n \ln[y_n(z)] - \ln(w_n!) \tag{26}$$

with

$$y_n(z) = \sum_k c_k \exp(j2\pi k x_n / d)$$

where the elements of $z = (c^T \; d)^T$ correspond to those of ζ, and those of $c = (c_0 \; c_1 \; c_{-1} \ldots c_K \; c_{-K})^T$ correspond to those of γ. Then, the complex gradient of Eq. (26) with respect to z is

$$-\sum_n \left(1 - \frac{w_n}{y_n(z)}\right) \frac{\partial y_n(z)}{\partial z} \tag{27}$$

Since the maximum of the likelihood function is, by definition, a stationary point, the $2K + 2$ elements of the maximum likelihood estimate $\tilde{z}$ of ζ must satisfy the $2K + 2$ nonlinear equations in z obtained by equating Eq. (27) to the null vector. The numerical solution for z representing the absolute maximum of the likelihood function described by Eq. (26) is the maximum likelihood estimate of the Fourier coefficients.

It follows from Eq. (25) that the expectations $y_n(\zeta)$ of the observations w_n are described by

$$y_n(\zeta) = X\gamma$$

where the n-th row of X is defined as

$$(1 \; \exp(j2\pi x_n/\delta) \; \exp(-j2\pi x_n/\delta) \ldots \exp(j2\pi K x_n/\delta) \; \exp(-j2\pi K x_n/\delta)$$

If the observations are normally distributed with covariance matrix W, then the maximum likelihood estimator of ζ has to minimize

$$[w - y(z)]^T W^{-1} [w - y(z)] \tag{28}$$

Hence, if only the Fourier coefficients are unknown, their maximum likelihood estimator is

$$\hat{c}_{W^{-1}} = (X^H W^{-1} X)^{-1} X^H W^{-1} w \tag{29}$$

For observations with a distribution different from normal, this estimator is no longer maximum likelihood but is still best linear unbiased. If, in addition, the period is unknown,

the estimation problem is recognized as a separable nonlinear least squares problem. The model is linear in the $2K + 1$ Fourier coefficients and nonlinear in the period d. This means that, in addition to Eq. (29), one further equation must be satisfied. This is the equation resulting from equating the derivative of the least squares criterion with respect to d to zero. If, in this equation, Eq. (29) is substituted for the Fourier coefficients c, a scalar nonlinear equation is obtained in the period d only. Hence, all that needs to be done is real, scalar root finding to estimate the period δ and substitute the estimate in the closed form of Eq. (29) for the Fourier coefficients. The estimates thus obtained are maximum likelihood if the observations are normal with covariance matrix W. They are weighted complex nonlinear least squares estimates with other error distributions.

If the covariance matrix W is unknown, the ordinary least squares estimator

$$\hat{c}_I = (X^{\mathrm{H}}X)^{-1}X^{\mathrm{H}}w \tag{30}$$

may be chosen, possibly combined with root finding for the period. This is a maximum likelihood estimator only if the w_n are independent and identically normally distributed about their expectations. For other distributions, it is a best linear unbiased estimator if only the Fourier coefficients are to be estimated, and the observations are uncorrelated with equal variance. In other cases, it is simply the ordinary least squares estimator.

A special case occurs if only the Fourier coefficients are to be estimated, the measurement points x_n are equidistant with interval Δ, the period is a known integer multiple of Δ, and an integer number of periods is observed. Under these conditions, the elements of $\hat{c}_I$ described by Eq. (30) may be shown to be equal to the DFT

$$\frac{1}{N}\sum_{n=1}^{N} w_n \exp[-j2\pi k(n-1)/M]$$

where $M\Delta$ is the period. Under the restrictive conditions mentioned, the DFT is, therefore, the maximum likelihood estimator if the observations are independent and identically normally distributed about their expectations. For other distributions, it is best linear unbiased if the observations are uncorrelated and have equal variance.

Measurement Errors and Resolution

Like precision and accuracy, resolution is a key notion in applied science and engineering. It is used in fields as diverse as radar, sonar, optics, electron optics, seismology and various forms of spectroscopy. An extensive review of resolution is presented in reference 10.

The most important form of resolution is two-component resolution.

Example 13. Rayleigh two-component resolution. As discussed in reference 10, Rayleigh considers observations described by

$$\alpha\{\mathrm{sinc}^2[2\pi(x-\beta_1)] + \mathrm{sinc}^2[2\pi(x-\beta_2)]\}$$

with $\mathrm{sinc}(x) = \sin(x)/x$. This is a pair of sinc-square components of equal height and located at β_1 and β_2, respectively. As the difference in location decreases, the components increasingly overlap and become increasingly difficult to distinguish visually. According to Rayleigh, the components are resolvable if the absolute difference of β_1 and β_2 exceeds 0.5. At this distance, the maximum of the one component coincides with the first zero of the other, and the component sum has two maxima and a relative minimum in between. Then, the ratio of the value at the relative minimum to that at the maxima may be shown to be 0.81. Later, this ratio has been generalized to other component functions, and the distance corresponding to this ratio has been called generalized Rayleigh resolution limit.

From this example, it is clear that this classical resolution limit and comparable ones proposed later are, in fact, measures of component width. Since in definitions such as Rayleigh's, the component functions are known, and the observations are exact, today, the model could be exactly fitted numerically to the observations with respect to the locations, the result would be exact, and there would in fact be no obvious limit to resolution. The reason why in practice unlimited resolution cannot be achieved is that observations exactly describable by two component functions do not occur. Therefore, it is not the distance of the components, but it is the errors in the observations, systematic and nonsystematic, that ultimately limit resolution. During the last decades, a number of measurement error based resolution limits have been proposed in the literature reviewed in Ref. 10. One of the most recent ones will now be described.

Suppose that a number of two-component observations $w = (w_1 \ldots w_N)^{\mathrm{T}}$ has been made, and that the two-component model

$$a[h(x;b_1) + h(x;b_2)] \tag{31}$$

is fitted to these observations with respect to a, b_1, and b_2. Then, depending on the set of observations available, two essentially different types of solutions for a, b_1, and b_2 may occur. In the first type, the solutions for b_1 and b_2 are distinct. This implies that the two components in Eq. (31) are resolved from the observations. In the second type of solution, the solutions for b_1 and b_2 exactly coincide. Then, the model corresponding to this solution is $2a\ h(x;b)$ with $b_1 = b_2 = b$. Thus, it is concluded that a one-component model is found as solution. This one-component solution is, of course, not found from exact two-component observations of the same functional family as the model fitted. However, it may result from error corrupted, two-component observations if the components seriously overlap.

At first sight, exactly coinciding solutions may look highly improbable. However, their coincidence is not caused by mere chance but by a structural change of the criterion of goodness of fit under the influence of the set of observations. In the N-dimensional Euclidean space of the observations, where the n-th coordinate axis corresponds to the n-th observation w_n, a set of observations is represented by a single point. If two-component models are fitted, this space may be divided into two parts. For observations in the one part, the criterion has an absolute minimum with $b_1 \neq b_2$. For observations in the other part, only a minimum can be shown to exist with $b_1 = b_2$. The boundary of both parts separates the sets of observations from which the components can be resolved from those

from which they cannot. Therefore, it is this boundary that constitutes the limit to resolution in terms of the observations. Of course, hypothetical, errorless two-component observations to which a two-component model of the same family is fitted are on the side of the boundary corresponding to resolution. However, nonsystematic and systematic measurement errors may move this point to the other side of the boundary, where resolution is impossible since the solutions coincide. Systematic measurement errors influence the location of the point around which sets of observations are distributed. This point represents the expectations of the observations. The systematic errors may move this point close to the boundary. The kind of distribution of the nonsystematic errors defines how the sets of observations are distributed around this point. Therefore, the probability of resolution is determined by both types of errors combined.

BIBLIOGRAPHY

1. C. Chatfield, *Statistics for Technology,* 3rd ed., London: Chapman and Hall, 1995.
2. A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics,* 3rd ed., Auckland: McGraw-Hill, 1987.
3. A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics—Vol.1 Distribution Theory,* London: Arnold, 1994.
4. A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics—Vol. 2 Classical Inference and Relationship,* London: Arnold, 1991.
5. Anonymous, *Guide to the Expression of Uncertainty in Measurement,* 1st ed., Geneva: International Organization for Standardization, 1993.
6. W. H. Press et al., *Numerical Recipes in Fortran; the Art of Scientific Computing,* 2nd ed., New York: Cambridge University Press, 1992.
7. A. Grace, *Optimization Toolbox for Use With MATLAB™, USCL's Guide,* South Natick, MA: The Math Works, 1990.
8. R. I. Jennrich, *An Introduction to Computational Statistics,* Englewood Cliffs, NJ: Prentice-Hall, 1995.
9. A. van den Bos, Complex gradient and Hessian, *IEE Proceedings Vision and Image Signal Processing,* **141** (6): 380–382, 1994.
10. A. J. den Dekker and A. van den Bos, Resolution—A survey, *J. Opt. Soc. Am.,* **14** (3): 547–557, 1997.

ADRIAAN VAN DEN BOS
Delft University of Technology