LIMITS AND ACCURACY IN MEASUREMENTS

G.A. Ososkov
Joint Institute for Nuclear Research, Dubna 141980, Russia

E.L. Kosarev
P.L. Kapitza Institute for Physical Problems, Moscow 119334, Russia

Keywords: Measurement, Sample, Accuracy, Statistics, Error, Bias, Estimation, Maximum likelihood, Least squares, Parametric method, Non-parametric, Resolution, Superresolution, Robust statistics

Contents

Summary

The natural way to find the limits of the accuracy in measurements is based on the probability theory and the mathematical statistics. In this general review distribution-free methods and non-parametric methods, point and interval estimations of the unknown parameters are discussed. The connection between maximum likelihood method and least squares method is shown. The special sections are devoted to robust approach and to resolution of digital signals. Cramér-Rao lower-bound for accuracy is demonstrated.

1. Introduction

The word **ACCURACY** (from Lat. accuratus – made with taking care of) has several definitions:
- the freedom from errors;
- property of a human statement to be close to truth;
- degree of conformity of a measurement to a true value, i.e. to a standard or to a model;

which are reflected by its synonyms: exactness, correctness, precision. Staying more in the EOLSS context we would rather consider the case of measurements distinguishing between **direct** and **indirect** measurements. Direct measurements are accomplished either by counting the number of some events within a given time interval (as for instance, for Geiger counter), or by comparing a measured object with a standard, i.e. its accuracy can be evaluated quantitatively in units of a minimal scale factor. However in contemporary sciences and technologies direct measurements are inherent in the lowest level of a procedure of more sophisticated indirect measuring of an observable phenomenon. Such phenomena are described, as a rule, by theoretical models with given quantitative characteristics of parameters. Thus indirect measurements suppose to be a subject of calculations that leads to the problem of the accuracy estimation from the set of measured values. This problem is caused not only by the complexity of a functional dependence connecting a chosen model, its parameters and measurements, but mainly, due to errors of the latter. These errors are inherent in any measurements, direct or indirect, regardless of the thoroughness, with which the measurements have been done. The

accuracy of parameters in question is inversely proportional to those errors. Therefore they have to be classified according to their sources and analyzed in order to be decreased as much as possible.

There are several types of errors distinguished depending either on their sources – such as **instrumental** and **model errors**, or on their statistical behaviour – such as **bias** and **random errors**.

Instrumental errors appear due to inevitable distortions introduced into the measurements by various maladjustments of a measuring device while its construction or by misalignments of its parts. Such errors can be observed in the process of a special calibration procedure, when an especially designed standard object is measured. Results of these calibration measurements are then handled to be compared with well known features of the standard. Such calibration data handling has twofold goals: (1) to evaluate and approximate distortions of the measuring device in order to compensate them mathematically; (2) to determine a functional transformation from the scales of measuring device to the standard coordinate system. From a mathematical point of view calibration problems belong to the more general class of unfolding problems described below.
Model errors are specific for hierarchical, indirect measurements and can often result in more serious errors in interpretation of experimental data. As soon as one tries to describe a certain phenomenon by a functional dependence on measured data and some parameters, then the choice of the type of function and values of its parameter can appear critical in verifying of such a description. We include here also errors of the method implementation, such as errors of approximation, round-off and discarding of expansion members which are of higher order of smallness. A typical example of such model errors appears when one tries to approximate observations of an unknown dependence by a polynomial. A wrong choice of this polynomial degree leads to an unavoidable approximation error. It is just the error of the wrong model and results usually in a significant accuracy loss.

However, all the errors listed above are developed in statistics of observations and, therefore, each of them can be classified statistically either as a bias (systematic error) or as a random error.

The systematic errors are caused by factors acting identically during the whole measuring process. The simplest example is weighing with wrong weights. It would always give you a wrong result unless you weigh a well-known standard weight, i.e. you make a calibration of your balance by calculating the difference between the previous biased measurement and *a priori* known weight of your standard. Then you can weigh any object and obtain its correct weight by adding that difference to the result of this biased weighing, which is, in fact, an example of the alignment transformation.

The random errors vary even for completely identical conditions of measurements depending on many occasional reasons which influences can not be taken into account in advance. We do not consider here gross errors of measurements that usually can be avoided by a careful experiment design or be eliminated later by a corresponding cut-off procedure.

Thus depending on the measurement process of any experiment some of errors listed above must be taken into account in order to improve the accuracy of measurements by a correct choice of statistical procedures embodying data handling algorithms.

2. Mathematical Formalism

In mathematical formulation we have a set of measurements (a sample)

$$x_1, x_2, \ldots, x_n \qquad (1)$$

to be processed statistically to extract the maximum of useful information related to the explored phenomenon with an acceptable level of accuracy. If our sample consists of equally distributed, independent random variables, then the first problem is usually to estimate their mean value and variance. More sophisticated problem is to estimate either the cumulative distribution function of our sample or its probability density function (p.d.f.). Depending on the nature of data and our *a priori* knowledge it can be done in several ways.

2.1. Distribution-free Methods

If the type of the sample distribution is unknown, one of **distribution-free methods** can be applied to estimate the sample mean value and even the distribution law of our sample. Theses methods are usually based on the **order statistics** $x_{(i)}$ obtained from (1) by reordering the sample in ascending order, so that $x_{(1)} \le x_{(2)} \le \ldots, \le x_{(n)}$ and the ordered measurements $x_{(i)}$ are called the order statistics. In particular, one of those statistics, namely $x_{(n/2)}$ named **the median** is a good estimation of the distribution mean value.

The *empirical probability distribution function* of the order statistics defined as

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ i/n & x_{(i)} \le x < x_{(i+1)} \\ 1 & x_{(n)} \le x \end{cases} \qquad (2)$$

can serve as a good estimation of the sample distribution law, whose accuracy is increased asymptotically with increasing n. That allows us to determine the type of the sample distribution by some of distribution-free **goodness-of-fit tests** in order to apply afterwards one of the parametric methods described in the next subsection.

2.2. Parameter Estimation

In a parametric case the type of the sample distribution is known and the problem is to estimate its parameters. Given the sample (1), estimation consists in determining either a value (so-called point estimation) or an interval most likely including the unknown parameter value in question (interval estimation).

2.2.1. Point Estimators

Both terms: *estimation* and *estimator* are often used. There is a minor difference between them: the first one often denotes the process or the procedure of the parameter estimation whereas the second one more often denotes the specific function of the sample data which is used for parameter estimation. We shall use both terms.

Estimators are constructed as functions of our sample data and, therefore, are random values, whose accuracy related properties can be expressed in probability terms only: its mean value, variance and probability of a big deviation from the estimated parameter. Thus having chosen an estimator, one can consider its goodness in terms of following basic properties:

- consistency,
- unbiasedness,
- efficiency,
- robustness.

An estimator is called *consistent* if its estimates converge toward the true value θ of the unknown parameter as the number n of measurements increases. The convergence is understood *in probability*, i.e. given any ε and any η, $\hat{\theta}_n$ is a consistent estimator of θ if an N exists such that

$$P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) < \eta$$

for all $n > N$. One of the most widely known estimators of the center of the sampling distribution is the *arithmetic mean* of the sample data (we call it further *the sample mean*)

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad (3)$$

Its consistency follows from the famous *law of large numbers* for the majority of distributions. Although as a function of random measurements the sample mean is a random variable, it is more precise than any of these measurements, since its variance is $\sqrt{n}$ times smaller.

It should be borne in mind that there are distributions for which the law of large numbers is not valid and the arithmetic mean is an inconsistent estimator for these distributions. As an example, consider the probability density function of the Cauchy distribution

$$p(x) = \frac{1}{\pi\left(1 + x^2\right)}, \quad -\infty < x < \infty. \qquad (4)$$

Both the mean value and the variance do not exist for the Cauchy distribution.

Denoting by E the *mathematical expectation* of a random variable we define the *bias* b of the estimator $\hat{\theta}_n$ as the deviation of its expectation from the true value θ_0,

$$b_n\left(\hat{\theta}_n\right) = E\left(\hat{\theta}_n - \theta_0\right).$$

Thus, an estimator is *unbiased* if for all n and θ_0

$$b_n\left(\hat{\theta}_n\right) = 0$$

or

$$E\left(\hat{\theta}_n\right)=\theta_0.$$

Let us take as an example such an important characteristic of any distribution as its variance, i.e. the expectation of squared deviations of a random variable from its mean

$$\sigma_x^2=E\left(x-E\left(x\right)\right)^2.$$

Calculating the *sample variance* one should replace the unknown value of the distribution mean by its statistical analog, i.e. the sample mean (3) that gives

$$S_x^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2, \qquad (5)$$

which is, in fact, an estimation of the variance of our sampling distribution. However the replacement we made introduces the bias in this estimator. As it is easy to calculate, its expectation is equal to

$$E\left(S_x^2\right)=\frac{n-1}{n}\sigma_x^2=\sigma_x^2-\frac{1}{n}\sigma_x^2,$$

which means that we have the bias term σ_x^2/n. The estimator (5) is consistent and the presence of the small bias is not important when the sample size n is very large, but for small n we have to correct our estimator to make it unbiased

$$\tilde{S}_x^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2.$$

The above two properties are important, but not enough to describe the goodness of an estimator. Since it is a random variable, its precision can be evaluated in terms of its variance. For instance, from the above mentioned consistent estimators of the distribution center, namely, the median and the arithmetic mean of the sample data, the second one has its variance smaller than for the median (in the majority cases, when the variance of the sampling distribution does exist). The arithmetic mean can be clearly considered as a more efficient estimator than the median.

Thus, in general, the *efficiency* of an estimator is determined by its variance: the smaller it is the more efficient is this estimator.

The estimator *robustness* means that it should be independent of the distribution, or insensitive to departure from the assumed distribution. In such a sense the median is more robust estimation for the sample distribution center than the sample mean (3), especially for distributions like Cauchy distribution (4), one having no mean value at all.

More detailed consideration of robust estimates is given in Section 3 below.

In the choice of a good estimator one can note a conflict between efficiency and robustness requirements. It is a typical situation when one wants to choose an estimator which must meet all requirements stated above and even some more needed to satisfy such realistic demands as minimum computer time or a simplicity in understanding and, in general, minimum loss of scientists' time. To find a compromise one must establish an order of importance between these requirements taking into account statistical and other merits, like cost or time (urgency of completing a research). In frames of our present considerations we focus ourselves further on statistical merits.

From this standpoint one of the most powerful statistical methods for estimating parameters is the *maximum likelihood method* (MLM) invented by R.Fisher (1912). Suppose for the sample (1) we know the probability density function $f(x,\Theta)$ common for each x_i with unknown parameter vector $\Theta = (\theta_1,\ldots,\theta_m)$. Then so-called **likelihood function**

$$L\left(\Theta\right) = \prod_i^n p\left(x_i\right). \tag{6}$$

is the density function for obtaining this sample if Θ is fixed. The MLM consists in finding an estimate of parameters $\hat{\Theta}$, which maximizes $L(\hat{\Theta})$. Since the maximum of L is also the maximum of $\ln L$, it is easier to maximize the latter function by solving *the likelihood equations*

$$\frac{\partial \ln(L(\hat{\Theta}))}{\partial \theta_k} = 0,\ k = 1,2,\ldots,m \tag{7}$$

in order to obtain *the maximum likelihood estimation* $\hat{\Theta}$. Its remarkable properties such as *asymptotic consistency*, *efficiency* and *normality* are proven.

Now one can easily find MLM-estimates for a known sample distribution.

Example 1. The Gaussian distribution is given by

$$f\left(x;a,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(x-a\right)^2}{2\sigma^2}\right) \tag{8}$$

For the sample (1) taken from a normal population one obtains

$$\ln\left(L\left(a,\sigma\right)\right) = -n\ln\left(\sigma\right) + \ln\left(2\pi^{n/2}\right) - \frac{1}{2\sigma^2}\sum_i \left(x_i - a\right)^2 \tag{9}$$

The solution of the likelihood equations gives two MLM-estimates

$$\hat{a} = \frac{1}{n}\sum_{i=1}^n x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left(x_i - \hat{a}\right)^2, \tag{10}$$

which we already had before.

Example 2. The Poisson distribution. It is a discrete distribution of the random variable taking values equal to whole positive numbers k with the probability

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{11}$$

For the sample $k_1,\ldots,k_n$ one has

$$\ln\left(L\left(\lambda\right)\right) = -\lambda n + \ln(\lambda)\sum_{i=1}^{n} k_i + \ln\left(\prod_{i=1}^{n} \frac{1}{\left(k_i\right)!}\right). \tag{12}$$

The solution of the likelihood equation gives the estimate of the parameter λ

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} k_i.$$

2.2.2. Interval Estimators

Each of the point estimators discussed above gives us a value intended to estimate an unknown parameter. As it was pointed out, these estimators are random by their nature, but obtaining a value we do not feel that its randomness is concealed and it could deceive about its accuracy or probability of being close enough to an unknown parameter. Therefore an experimenter prefers to use the estimators that include explicitly the range

$$\theta_a \le \theta \le \theta_b,$$

which contains the true value θ_0 with probability β. Given a measurement x from a p.d.f. $f\left(x|\theta\right)$ with a known parameter θ, the probability content β can be calculated as

$$\beta = P\left(a \le x \le b\right) = \int_a^b f\left(x|\theta\right)dx. \tag{13}$$

However in our case we have an unknown parameter and too large arbitrariness in choosing the interval borders a and b. It would be better to choose an interval which has minimal length among all intervals $\left[\theta_a,\theta_b\right]$ with the same probability β. Such intervals are called *confidence intervals* for θ *with probability* β. Since the parameter θ is unknown, one has to take a different variable $z = z(x,\theta)$, a function of the measurement x and the parameter θ, but such that its p.d.f. is independent of the unknown θ. If it can be found, we can re-express Eq.(13) as a problem of interval estimation: given β, find the optimal range $\left[\theta_a,\theta_b\right]$ in θ-space such that

$$P\left(\theta_a \le \theta_0 \le \theta_b\right) = \beta. \tag{14}$$

It is better to explain this scheme with a particular example of the confidence interval for the mean of normally distributed sample data with the cumulative distribution function

$$\Phi\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt. \tag{15}$$

When both distribution parameters μ and σ are known one can calculate β from (13):

$$\beta = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

However when μ is unknown (but σ is known), one can instead calculate the probability β that some of functions of our measurements, say the sample mean $\overline{x}$, lie in an interval that includes its unknown mean. Let us take a symmetrical interval $[\mu - c, \mu + c]$. Then

$$\beta = P\left(\mu - c \le \overline{x} \le \mu + c\right) = \frac{1}{\sigma^*\sqrt{2\pi}} \int_{\mu-c}^{\mu+c} \exp\left(-\frac{\left(t-\mu\right)^2}{2\sigma^{*2}}\right) dt = \Phi\left(-\frac{c}{\sigma^*}\right) - \Phi\left(\frac{c}{\sigma^*}\right), \tag{16}$$

where for the sample mean $\sigma^* = \sigma/\sqrt{n}$. We can now invert the probability statement in (16) in order to take the form of the statement (14):

$$\beta = P\left(\overline{x} - c \le \mu \le \overline{x} + c\right).$$

As it is known in the case of the normal distribution, one can obtain $\beta = 0.95$ if the constant c is chosen as

$$c = 1.96\,\sigma^* = 1.96\,\sigma/\sqrt{n}.$$

2.3. Evaluation of Dependencies

Now we consider the more complicated set of two-dimensional observations

$$\left(x_1, y_1\right), \ldots, \left(x_n, y_n\right) \tag{17}$$

which we want to describe by a known function f embodying our model assumption by means of parameters in question $\theta_1, \theta_2, \ldots, \theta_m$. It gives us the following system of equations

$$y_i + e_i = f\left(x_i; \theta_1, \theta_2, \ldots, \theta_m\right), \quad i = 1, 2, \ldots, n, \tag{18}$$

where e_i are the errors of measurements, which are supposed to be random variables all with zero mean values and common distribution function $F_e(x)$; and n is a number of measured points.

For the easiest 2D-case Eq. (18) can be considered as the set of residuals between the measured points and a curve to be fitted to them by varying the vector $\overline{\Theta}(\theta_1,\theta_2,\ldots,\theta_m)$:

$$e_i = y_i - f\left(x_i;\overline{\Theta}\right), i = 1,2,\ldots,n. \tag{19}$$

where x_i is a given abscissa of the *i*-th measurement y_i.

2.3.1. Least Squares and Maximum Likelihood

The least squares method (LSM) was independently invented by C.F.Gauss (1805) and A.Legendre (1809) and it consists in **minimizing the sum of squares of those residuals** (19)

$$S\left(\overline{\Theta}\right) = \sum_{i=1}^{n} (y_i - f(x_i;\overline{\Theta}))^2 \tag{20}$$

with respect to unknown parameters. Since $S(\overline{\Theta})$ is a quadratic function of its arguments, its minimum is reached when $\partial S/\partial\theta_k = 0; k = 1,2,\ldots,m$. This system of equations is especially simple in the frequently used case of the linear regression:

$$f\left(x_i;\overline{\Theta}\right) = \sum_{j=1}^{m} \phi_j\left(x_i\right)\cdot\theta_j + e_i, \quad i = 1,\ldots,n, \tag{21}$$

where $\phi_j(\cdot)$ is known set of m linearly independent basic functions (e.g.$1, x, x^2,\ldots,x^m$); e_i - an accidental measurement errors; θ_j are as above unknown regression parameters $(j = 1,\ldots,m)$ which should be estimated by use of our data sample. Then the equation system for parameter estimations becomes linear. It is called **the normal system of equations**.

Optimal properties of LSM curve fitting follow from the fact that LSM is a particular case of the more general maximum likelihood method. As in Section 2.2.1 under the assumption that all residuals (19) are independent random variables with zero mean and a common p.d.f. $p(e)$ one has the probability of occurrence of our particular sample $e_1, e_2,\ldots,e_n$, is proportional to likelihood function

$$L(\overline{\Theta}) = \prod_{i}^{n} p\left(e_i\right). \tag{22}$$

Then we choose as the likeliest estimation of parameters the value of $\overline{\Theta}$ in which (22) has its maximum. The solution of the likelihood equation

$$\frac{\partial \ln(L(\bar{\Theta}))}{\partial \theta_k} = 0,\ k = 1,2,\ldots,m \qquad (23)$$

gives the likelihood estimate $\hat{\bar{\Theta}}$. Under the crucial assumption of normality of the p.d.f. in (22), i.e.

$$p\left(e_i\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right), \qquad (24)$$

one immediately has that MLM reduces to LSM. Since

$$\ln(L(\bar{\Theta})) = \ln\left(\prod_{i=1}^{n} p\left(e_i\right)\right) = \sum_{i=1}^{n} p\left(e_i\right),$$

the logarithmic likelihood function becomes equal to

$$-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{e_i^2}{\sigma^2}\right) + \text{const},$$

which has its maximum exactly at the same point where (20) has its minimum. This implies that all above-mentioned good properties of a LSM-estimation are valid only if the normality assumption holds.

2.3.2. Measurements with Different Accuracy

Again we assume that the errors of measurements are distributed normally about zero, i.e.

$$y_i = f\left(x_i;\bar{\Theta}\right) + e_i,\ E\left(e_i\right) = 0,\ E\left(e_i^2\right) = \sigma^2 = 1/w_i.$$

The LSM requires

$$S = \sum_{i=1}^{n} w_i e_i^2 \Rightarrow \mathbf{min}. \qquad (25)$$

The terms in the sum of squares are now *weighted* by the reciprocals of the variances. We explain details of S minimization for the linear case (21) with $\phi_j(x_i) = x_i^{j-1}$ in the following notations

$$\mathcal{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots, & x_1^m \\ 1 & x_2 & x_2^2 & \ldots, & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots, & x_n^m \end{pmatrix}, \ \mathcal{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \Theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_m \end{pmatrix}.$$

Defining also the weight matrix

$$\mathbf{W} = \begin{pmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_n \end{pmatrix}$$

we can rewrite (25) in its matrix form as

$$L = \boldsymbol{\mathcal{E}}^T \mathbf{W} \boldsymbol{\mathcal{E}},$$

where $\boldsymbol{\mathcal{E}} = \mathcal{Y} - \mathcal{X}\Theta$. Then the solution of the normal system of equations becomes

$$\hat{\Theta} = \left(\mathcal{X}^T \mathbf{W} \mathcal{X} \right)^{-1} \mathcal{X}^T \mathbf{W} \mathcal{Y}$$

It gives us estimations of all $m+1$ parameters of our fit. Their covariance matrix is

$$\operatorname{cov}\left(\hat{\Theta}\right) = \mathcal{C} = \left(\mathcal{X}^T \mathbf{W} \mathcal{X} \right)^{-1}.$$

It contains squared errors of parameters on its diagonal, but since those parameters are correlated, one can obtain also covariances of *i*-th and *j*-th parameters as the element c_{ij} of this matrix.

3. Robust Approach

Actually, the crucial normality assumption is very often violated due to contamination of the measurements by points of noise or background. Then due to quadratic view of the functional S in (20) the contribution to it by any sample point-outlier could seriously disturb the estimation of parameters that causes a significant loss of accuracy. To calculate correct values of parameters one should use only the measurements from a close vicinity of the fitted function $f\left(x;\overline{\Theta}\right)$. All others should have much less impact or be completely negligible. This idea can be implemented by attributing **special weights** to each of measurements. Values of these weights must decrease with the growth of residuals e_i, i.e. with the growth of the distance from the fitted curve. This approach named **robust**, i.e. resistive to contamination, has been proposed by P.Huber (1972). Let us take as an example a linear regression dependence (21) to expound how to find the best way to process the noisy data by such a robust approach.

We describe the contaminated distribution of measurement errors e_i by J.Tukey's *gross-error model*

$$f(e)=(1-\beta)\cdot p(e)+\beta\cdot h(e), \tag{26}$$

where β is a parameter of contamination, $p=N(0,\sigma_m^2)$ is the Gauss distribution (24), and h is some long-tailed noise distribution. Using the maximum likelihood method

$$L=\prod_{i=1}^{N} f(e_i)\Rightarrow \mathbf{max}$$

we obtain the following system of non-linear equations

$$\sum_{i=1}^{N} w_i\cdot\left(Y_i-\sum_{j'}^{p}\phi_{j'}(x_i)\cdot\theta_{j'}\right)\cdot\phi_j(x_i)=0 \tag{27}$$

with some optimal weights w_i depending on relations of p and h distributions

$$w_i=\frac{p(e_i)}{p(e_i)+\frac{\beta}{1-\beta}\cdot h(e_i)}. \tag{28}$$

Taking $p(e)$ from (24) and $h(e)$ to be uniform and equals $h=h_0$ for distinctness one obtains the optimal weight function as

$$w_{opt}(e)=\frac{1+c}{1+c\cdot\exp\left(\frac{e^2}{2}\right)}, \tag{29}$$

with the constant

$$c=\frac{h_0\beta\sigma\sqrt{2\pi}}{1-\beta}.$$

These weights are non-linear functions of the parameters in question. Therefore an iterative procedure was elaborated, in which the weights are recalculated at each iteration in accordance with the new values of parameters. This procedure is referred as re-weighted LSM (RWLSM). A polynomial expansion of these optimal weights up to the fourth order leads to the approximation.

$$w(t)=\begin{cases}\left(1-(t/c_T)^2\right)^2, & \text{if } |t|<c_T\\ 0, & \text{otherwise.}\end{cases} \tag{30}$$

We obtain, in fact, the famous Tukey's bi-weights which are easier to calculate than the optimal ones. It is also recommended to choose the cutting parameter $c_T = (3 \div 4)\sigma$. If there is no *a priori* information one can initiate the iterations with $w_i^{(0)} = 1$.

The parameter σ should be also recalculated at each iteration

$$\sigma^{(k)} = \sqrt{\frac{\sum_i w_i^{(k-1)} \left(e_i^{(k-1)} \right)^2}{\sum_i w_i^{(k-1)}}} \qquad (31)$$

Following P.Huber one can consider this procedure as descending M-estimate. Thus the robust approach allows us to handle data accurately even in presence of contamination, but if one is sure that data to be processed are free from background, noise or other contaminating factors, then there is no need to use any RWLSM-procedure, since the LSM is proven to be the most efficient method in this case.

4. Resolution of Digitized Signals

It should be noted, however, at this point that the situation with real measurements in majority of contemporary physical, chemical or biological experiments is much far from the idealized models described above. Almost any modern measuring device is a quite sophisticated physical apparatus, which registers results of measurements in a discrete form after a digital procedure. In order to improve the resolution and accuracy of such detectors, they are designed as granular structures consisting of an array (raster) of cells (pads). Therefore if a signal to be detected looks like a single thin peak after being registered it would be smeared by the measuring device between several adjacent cells and then it is discretized in a view of a histogram, whose dimension depends on experimental data. An accidental measurement error is added to each histogram bin during registration. Additionally, background noise gives also its contribution to every cell of our device raster.

The basic problem is: to reconstruct the original signal position and its other parameters (its amplitude or the volume under its surface, its half width etc) from the registered histogram. Depending on its formulation this problem can be solved in either non-parametric or parametric ways.

4.1. Non-parametric Approach

The first non-parametric approach also named *the unfolding problem* is applied when the parameterization of the problem is unknown. Taking one-dimensional case for simplicity one can present a measurement result as one-dimensional histograms $\{F_i\}$ for $i = 1, 2, \ldots, n$ whose bin sizes and centers we denote respectively as Δ_i and x_i. These measured values are distorted by the detector influence and spoiled by noise. We assume that the values of the random function $F_i = F(x_i)$ are independent at different x_i and they do not contain systematical errors. Let us denote by $K(x, y)$ the

instrumental or the point spread function of our detector. In most cases it is symmetrical and depends only on the difference of its arguments, i.e. $K(x, y) = K(x - y)$ as, for instance, a Gaussian

$$K\left(s\right) = \exp\left(-\frac{s^2}{2\sigma_s{}^2}\right)$$

with σ_s usually supposed to be a known constant. Then one can, in principle, reconstruct the original signal in question $f(x)$ as a solution of the system of following integral equations of the first kind

$$F_i = \int_{\Delta_i} K\left(x-y\right) f\left(y\right) dy = \int_{x_i-\Delta/2}^{x_i+\Delta/2} K\left(x-y\right) f\left(y\right) dy,\ i = 1,2,\ldots,n. \tag{32}$$

The finite size Δ of the detector granularity can be expressed as the function $h(x)$

$$h(x) = \begin{cases} 1,\ |x| \le \Delta/2 \\ 0,\ |x| > \Delta/2. \end{cases} \tag{33}$$

It allows us to take into account the signal distortions due to the histogramming process of the experimental data storing. Its influence can be expressed mathematically by convolution of the point spread function $K(s)$ with the granularity function $h(x)$ giving the modified instrumental function

$$K_1\left(s\right) = \int h\left(s-r\right) K\left(r\right) dr. \tag{34}$$

Therefore the system (32) is transformed to

$$F_i = F\left(x_i\right) = \int K_1\left(x_i - x\right) f\left(x\right) dx. \tag{35}$$

Due to discreteness of our model the integral equation system (35) is reduced to the system of algebraic equations suitable for implementing by computer

$$\sum_{j=1}^{m} P_{ij} G_j = F_i,\ i = 1,2,\ldots,n, \tag{36}$$

where $G_j = f(y_j)$ are unknown variables and

$$P_{ij} = K_1\left(x_i - y_j\right) \cdot \left(y_{j+1} - y_j\right). \tag{37}$$

Unfortunately in most cases the system (36) is ill-posed, i.e. it has no solution at all since its right-hand side is a random vector and the system matrix P_{ij} is practically degenerated. However as it can be

proved, when one looks for a solution on a compact set, the problem can be solved for a sufficiently wide class of kernels including the convolution $K_1(s)$.

In the case of Gaussian distribution function of input data the logarithmic likelihood function may be written as

$$L = -\frac{1}{2}\sum_{i=1}^{n}\frac{\left(F_i - S_i\right)^2}{\sigma_i^2} + \text{const}, \qquad (38)$$

where σ_i^2 are the noise variances at the *i*-th experimental point and the values of S_i are defined by the formula

$$S_i = \sum_{k=1}^{m} P_{ik} G_k, \; i = 1, 2, \ldots, n.$$

The sought maximum of the likelihood function L can be calculated by means of an iterative procedure. As it has been proved by E.L.Kosarev (1993), this procedure converges to the maximum of the likelihood function (38).

E.L.Kosarev has also shown that the resolution obtained reaches the possible theoretical limit. Such a limit for close signals follows from the well-known Shannon's theorem for the maximum speed of the data transmission via the channel having a noise. The **resolution** of any linear device with the instrumental function $K(x)$ can be defined as the effective width of this function, i.e.

$$D = \int_{-\infty}^{\infty} K^2\left(x\right) dx, \qquad (39)$$

providing the normalizing condition at the origin $K(0) = 1$.

The resolution of spectral devices can be improved in comparison to D using the modern techniques for solving integral equations, thus **superresolution** is achieved. We define **the superresolution factor** as the ratio of D to the separation δ between two narrow lines which can be recovered after the deconvolution procedure

$$SR = D/\delta. \qquad (40)$$

We should remember that according to the well-known Rayleigh's definition of resolution $\delta = D$, so in this case the superresolution factor is $SR = 1$. When the resolution is improved mathematically, $SR > 1$. The improvement is always limited by noise. At zero noise an exact solution of Eq.(35) can be found, which corresponds to an infinite superresolution.

The highest possible superresolution factor is closely related to the Shannon theorem on the highest possible transmission rate of information through a noisy channel. When a spectrum is not parametric, i.e. the function we sought for cannot be described by a simple formula with a few parameters, the limiting superresolution factor is

$$SR = \frac{1}{3}\log_2\left(1 + E_s/E_n\right). \qquad (41)$$

Here E_s and E_n

$$E_s = \int_{-\infty}^{\infty} F^2(x)dx, \;\; E_n = n\sigma^2,$$

are the signal energy and the noise one. Here n is the number of experimental data points, and σ^2 is the variance of input noise. If the signal-to-noise ratio is expressed in decibels $dB = 10\log\left(E_s/E_n\right)$, the approximate expression for the superresolution limit is

$$SR \simeq dB/10. \qquad (42)$$

Using the computer program package RECOVERY cited in bibliography one can reach the Shannon superresolution limit.

4.2. Parametric Approach

The second parametric approach is well-known as parameter fitting. It is obvious that parametric methods must be more accurate than non-parametric ones, since the parameterization itself brings an essential information related to processed signals. Even such a general knowledge as the signal symmetry is enough to apply the easy-to-calculate center of gravity (COG) method for estimation of the signal centroid:

$$x_{cog} = \frac{\sum_i a_{i,j}\overline{x}_i}{\sum_i a_{i,j}} \quad y_{cog} = \frac{\sum_j a_{i,j}\overline{y}_j}{\sum_j a_{i,j}}, \qquad (43)$$

where $a_{i,j}$ is 2D histogram presenting a detector response to the current signal, $\overline{x}_i, \overline{y}_j$ are the middle points of the corresponding bins. The high speed and universality of this method made it the most popular for the majority of discrete detectors, unless the signal overlapping due to the high occupancy in many of modern experimental systems. It occurs when the probability of two and more signals to be overlapped is high enough.

Therefore more elaborated methods based on LSM robust modifications are developed by G.Ososkov (2000). Due to the problem of non-linearity they are iterative, so the COG solution can be used as an initial approximation. A multiparametric model of a signal obtained as a superposition of two Gaussian peaks with different positions and amplitudes is used. Corresponding problem of non-linear LSM functional minimization is solved there by a paraboloidal approximation method. As one more example of successful applications to the same problem of the close signal resolution, a method using wavelet transforms can be also pointed out. There are a great number of such examples of successful solutions of signal recovery parametric problems in different fields. All these methods have various degree of

accuracy depending on the noise or contamination level, but from statistical point of view all they intend, as the matter of facts, to provide various estimations of parameters.

Thus a question arises: is there a limiting accuracy and how to reach it?

4.3. Cramér-Rao Lower-bound for Accuracy

For the sake of simplicity let us consider a model of estimating of the single signal location parameter x_0 from a sample of n measurements:

$$y_i = A\psi\left(x_i - x_0\right) + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

Here $\psi(x)$ is the signal shape, A is its amplitude, and a random noise ε_i has the Gaussian distribution $N(0,\sigma)$, i.e.

$$p\left(\varepsilon_i\right)d\varepsilon_i = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{\varepsilon_i^2}{2\sigma^2}\right]d\varepsilon_i\,,$$

where different noise sample units are assumed to be (for a simplicity again) independent and non-correlated: $\overline{\varepsilon_i\varepsilon_j} = \sigma^2\delta_{ij}$. Applying the standard maximum likelihood method for estimating the signal amplitude A and location parameter x_0 one has to maximize the likelihood function of the sample

$$L = \prod_{i=1}^{n} p\left(\varepsilon_i\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2\right] =$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - A\psi\left(x_i - x_0\right)\right]^2\right].$$

It is reduced to maximization of the logarithmic likelihood function

$$l = \ln L = \text{const} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - A\psi\left(x_i - x_0\right)\right]^2.$$

According to the Cramér-Rao inequality *for unbiased estimates* one has

$$D\left(x_0\right) \geq \frac{1}{E\left[\left(\frac{\partial l}{\partial x_0}\right)^2\right]}.$$

Here, as usual, the symbol E denotes the mathematical expectation obtained by various random realizations of ε_i for $i = 1, 2, \ldots, n$. Taking the partial derivatives with respect to the parameter x_0 one

obtains the following formula for the Cramér-Rao lower-bound for the accuracy of the signal location parameter x_0 under condition of the negligible correlation between x_0 and A

$$D\left(x_0\right) \geq \frac{\sigma^2}{A^2} \cdot \frac{1}{\sum_{i=1}^{n}\left[\frac{\partial \psi\left(x_i - x_0\right)}{\partial x_0}\right]^2}. \qquad (44)$$

This final formula is valid for arbitrary shapes of the signal $\psi(x)$ and arbitrary ratio between the characteristic scale D of the function $\psi(x)$ and the bin size Δ (the latter can be treated as the distance between measured points).

Let us consider two limiting cases. In the first one when $\Delta \ll D,$ the Eq.(44) can be approximated by substituting of the sum in the denominator by the corresponding integral

$$\Delta \cdot \sum_{i=1}^{n}\left[\frac{\partial \psi\left(x_i - x_0\right)}{\partial x_0}\right]^2 \approx \int_{-\infty}^{\infty}\left(\frac{\partial \psi}{\partial x}\right)^2 dx. \qquad (45)$$

Here Δ denotes the histogram bin size. Computing the corresponding integral for the signal shape $\psi(x) = \exp\left(-\left(x/D\right)^2\right)$ one obtains eventually the formula

$$D\left(x_0\right) \geq \frac{\sigma^2}{A^2} \cdot \frac{1}{\sum_{i=1}^{n}\left[\frac{\partial \psi_i}{\partial x_0}\right]^2} \approx \frac{\sigma^2}{A^2} \cdot \frac{D\Delta}{\sqrt{\frac{\pi}{2}}}$$

and its approximate estimation

$$\delta x_0 = \sqrt{D\left(x_0\right)} \approx \frac{\sigma}{A} \cdot \frac{\sqrt{D\Delta}}{\left(\pi/2\right)^{1/4}} \approx 0.89\frac{\sigma}{A}\sqrt{D\Delta}. \qquad (46)$$

If $\sigma/A = 5\%, D = 4, \Delta = 1,$ one has $\delta x_0 \approx 0.1,$ i.e. 10 times better than bin size.

In the second limiting case both characteristics scales D and Δ are approximately of the same size

$D \sim \Delta.$

In this case the approximation Eq.(45) is not valid and one should compute explicitly the sum in the denominator of the Eq.(44).

Acknowledgments

The authors are grateful to Prof. Yu.Tsipenyuk for the opportunity to contribute to the EOLSS and for the help suggestions in the preparation of the manuscript.

Glossary

Bias $b(\theta)$ in statistical terminology is determined as the deviation of the mathematical expectation E of some estimation T_θ of a parameter θ from the true value of this parameter $b\left(\theta\right)=E\left(T_\theta\right)-\theta.$ If $b(\theta)\equiv 0$ then the estimation T_θ is **unbiased**.

Distribution function (often: cumulative distribution) $F(t)$ of some random variable **x** is determined as the probability $\mathcal{P}$ of the event that $\mathbf{x}<t\quad F(t)=\mathcal{P}\left(\mathbf{x}<t\right),-\infty<t<\infty.$ The function $F(t)$ is non-decreasing and continuous from the left.

Error ε of a measurement x of some value x_0 is usually determined as deviation of the result of the measurement x from the value of x_0, $\varepsilon=x-x_0.$ There are two kinds of errors: random and bias and both are essential as a contribution to the final accuracy of measurements. There is a lower limit of any accuracy achievable by data handling based on the Cramér-Rao theorem.

Estimation is in a general case a function f of measurements $x_1,\ldots,x_n$ and it is therefore a random variable. The value of this function $f(x_1,\ldots,x_n)$ is *an estimation* of unknown characteristic – the latter one can be a set of parameters or a distribution function.

Least squares is a parameter estimation method which consists of minimizing the sum of squared deviations of the data $y_i,i=1,\ldots,n$ of the measurements from some function $y(x;a_1,\ldots,a_m)$ of *x*-coordinate to be fitted by a proper choice of unknown parameters $a_1,\ldots,a_m$

$S=\sum_{i=1}^{n}\left(y_i-y\left(x_i;a_1,\ldots,a_m\right)\right)^2\Rightarrow\mathbf{min}.$ If measurements $y_i,i=1,\ldots,n$ have the Gauss distribution the Least Squares Method (LSM) is a particular case of the **Maximum Likelihood Method** (MLM).

Likelihood function $L\left(t\middle|\theta\right)=L\left(t_1,\ldots t_n\middle|\theta_1,\ldots,\theta_m\right)$ is determined as the probability that measurements $x_1,\ldots,x_n$ are taken fixed values $t_1,\ldots,t_n.$ This definition is valid only for discrete variables $x_i.$ In continual case one must replace 'probability' to 'probability density'. According to the MLM one can find the values of unknown parameters $\theta_1,\ldots,\theta_m$ by maximization of the likelihood function in the *m*-dimensional parameter space $\{\theta_k\},\ k=1,\ldots,m.$ The arguments of the likelihood function $t_i=x_i,\ i=1,2,\ldots,n$ are considered as fixed while the maximization process.

Parametric method is in fact the MLM when one looks for some unknown parameters $\theta_1,\ldots,\theta_m$ from measurements $x_1,\ldots,x_n$ under the condition that number of parameters $m\ll n.$ If the last inequality does not hold, the parametric method can sometimes result in inconsistent estimations.

Resolution (sometimes denoted as **resolving power**) is determined for any linear registering device (e.g. a spectrometer in optics) with the instrumental function $K(x)$ as

$$D=\int_{-\infty}^{\infty}K^2\left(x\right)dx,$$

providing the normalizing condition $K(x_0)=1$ where x_0 is the abscissa point where $K(x_0)=\max$. The resolution of a registering device can be improved in comparison with D by applying the modern techniques for data treatment, thus **superresolution** can be achieved.

Robust approach is a way of statistic estimation design to be less sensitive to disturbance of the basic hypotheses concerning the data and contaminating noise.

Sample of the size n is a set of elements $x_i,\ i=1,\ldots,n$ randomly chosen from a general population, i.e. another set elements $\{\xi\}$ of infinite size with the distribution function $F(\xi)$. According to this definition elements $x_i,\ i=1,\ldots,n$ are equally distributed, independent random variables. It should be mentioned that experimentalists prefer usually to say: "data" instead of "sample", "measure" instead of "draw a sample", and "observable space" instead of "general population".

Statistics in context of this chapter is the one of two connected parts of the mathematical science "Probability theory and mathematical statistics". Mathematical statistics is more 'experimental' science with comparison of a probability theory, because we have to draw information from the experimental data. We can consider problems of both theories as direct and inverse.

Bibliography

Cramer H. (1999) *Mathematical Methods of Statistics*, Princeton Univ. Press, 575p., Princeton, N-Y. [This book presents all essential aspects and approaches to mathematical statistics in a general and comprehensive way]

Van der Waerden B.L. (1969) *Mathematical Statistics*, 367p., Springer Verlag. [This is a widely used textbook that presents essential principles of mathematical statistics]

Brandt S. (1999) *Data Analysis. Statistical and Computational Methods for Scientists and Engineers.* 3-rd Edition, 652 p., Springer Verlag, .,. [It is a well known textbook where a good collection of recipes of statistical data analysis is illustrated by many Fortran programs]

Huber P. (1981) *Robust Statistics*, J. Willey & Sons, 308p., NY. [This book presents main approaches to robust methods of statistics]

Mosteller F., Tukey W. (1977). *Data analysis and regression: a second course in statistics*, 588p., Addison-Wesley. [The book of one of the most experienced specialists in exploratory data analysis with many practical recommendations and robust algorithms]

Ososkov G.A. (1987). *Robust regression for the heavy contaminated sample*, Proc. 2nd International Tampere Conference in Statistics (Tampere, Finland) [At this paper optimal weight are proposed for the robust estimates of regression parameters, joint estimates of location and scale parameters are also studied]

Shannon C.E. (1948) *A mathematical theory of communication*, Bell Sys. Techn. Journal, **27**, pp.379-423, 623-656. See also in the URL: http://www.cs.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf. [The introduction to the communication theory, where the famous theorem about limitations of signal transfer via a noisy channel is proven]

Shannon C.E. (1949) *Communication in the presence of noise*, Proc. of the IRE, **37**, No.1, pp.10-21. [Another classical C.Shannon paper, where his theorem is presented using geometric approach]

Kolganova E., Kosarev E., Ososkov G. (2000) *Superresolution algorithms for data analysis of discrete detectors in nuclear physics*, Nucl.Instr.Meth. **A443** No.2-3, pp.464-477. [In this survey parametric and non-parametric approaches are presented for the recovery of particle coordinates from distorted and noisy data with limiting superresolution efficiency]

Kosarev E.L. et al (1993) *Programs for signal recovery from noisy data using the maximum likelihood principle. Part 1: General description*, Computer Physics Communications, Vol.74, No.3, pp.335-348, *Part 2: Program implementation.* ibid, pp.349-357. [Theory of a robust method for the signal recovery and description of the RECOVERY program, implementing this method]

Biographical Sketches

Ososkov, Gennadii Alexeevich, graduated from the Moscow State University. in 1953, becoming candidate of sciences in 1957, and doctor of physico-mathematical sciences in 1987. From 1961 until the present time he has worked at Joint Institute for Nuclear research, Dubna Russia, now are being the leading scientist of this Institute. In addition he is a professor of computational statistics of the Dubna University and Ivanovo State University. His scientific interests include: mathematics, theory of probability, mathematical and computational statistics, pattern recognition, artificial intelligence, Monte-Carlo simulations. As an experienced expert in robust data handling in experimental physics, G.A.O. has been many times invited for scientific consultings to famous physical centers (to the Max Planck Institute for Nuclear Physics, Heidelberg, Germany in 1995, 1996, 1997, to the Yale University, New Haven, and Alabama University, Birmingham, USA in 1998, to DESY, Hamburg in 2000). In 1997 he was awarded by the first JINR prize, in 2000 the presidium RAS has awarded him by the monthly state grant in the field of information science for three years. G.A.O. has published more than 200 papers in scientific journals and proceedings of international conferences. He is the author of the monograph: "Pattern Recognition Application in High Energy Physics ", Serie "Lectures for young scientists", No.26, JINR publ. P10-83-187, Dubna, 1983, and coauthor of the second one: "Automation of measurements and data processing of physical experiment data ", Moscow State Univ. Publ., Moscow, 1986 (both in Russian)

Kosarev, Evgenij Leonidovich, graduated from the Moscow Institute for Physics and Technology (MIPT) in 1962, becoming candidate of sciences in 1971, and doctor of physics and mathematical sciences in 1995. From 1962 until the present time he has worked at P.L.Kapitza Institute for Physical Problems, Moscow, Russia, now are being the head scientist of this Institute and Head for computing there. He is also an Associate Professor in MIPT. His scientific interests include: experimental physics inverse problems for applications in X-ray detectors, any kind of spectroscopy, high energy physics particle detectors, muon-spin-rotation experiments, etc…He was invited for research visits to world physics laboratories: Daresbury Laboratory, UK in 1980, 1988, CERN, Switzerland in 1982, Photon Factory, KEK, Japan in 1990. E.L.Kosarev has published more than 50 papers in scientific journals and proceedings of international conferences.

