Measurement 41 (2008) 1100–1104

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Fully statistical approach for regression analysis

Petre Cătălin Logofătu *, Dan Apostol

National Institute for Laser, Plasma and Radiation Physics, Laser Department, P.O. Box MG-36, Măgurele, Romania

ARTICLE INFO

Article history:
Received 11 December 2006
Received in revised form 8 March 2008
Accepted 10 March 2008
Available online 15 March 2008

Keywords:
Error analysis
Measurement theory

ABSTRACT

A new, fully statistical approach for regression analysis is presented and used for deriving the formula for the estimation error of the parameters of the fit and the associated joint confidence levels assuming a normal (Gaussian) distribution of the measurement errors and using a type A evaluation of the uncertainties. The key feature of the approach consists in two complementary parameterizations of the error space that are equivalent to a change of coordinates. This feature makes possible all the derivations and gives a marked statistical character to the approach. Although this approach is more lengthy and laborious than the usual one, it has the advantage that follows step by step all the intricacies, statistical and topological, of the regression analysis, and the final formulae do not appear as black boxes to be used such as they are, but all their components have established meanings.

1. Introduction

In metrology, besides the actual measured value of the quantities and the estimated value of the parameters, one needs to know their uncertainty. One common situation is to fit data to a nonlinear function using the least squares method (the more simple case of linear functions can be inferred from the formulae for the nonlinear function as a particularization). The case is amply analyzed in the scientific literature [1,2]. We think, however, there are some new contributions that we can bring to the subject. Namely we show a fully statistical approach in the derivation of the error analysis formulae for the case when a Gaussian distribution of error is assumed and a type A evaluation of uncertainties is used. Although our formulae are basically the same as those shown in Bevington's book [1], they are obtained in a different way. The difference between the approach described by Bevington and ours is similar with the difference between the phenomenological and the statistical approach in thermodynamics.

The confidence levels and the limits for the parameters fit values are calculated using $\Delta\chi^2$ isocontours (see Fig. 1). The probability for a certain value of $\Delta\chi^2$ is not the same as the probability for a certain parameter estimation error, but it is a good approximation and mathematically more simple to calculate. The technique for error analysis presented here is a negative, "pessimist" approximation. It assumes the worst situation, and in some cases the real error is smaller.

2. Sets of measurement errors

A Gaussian distribution of the measurement errors is assumed. Although our analysis is meant for the fit to nonlinear functions, in all the derivations we kept only the first order terms of the Taylor expansion for simplicity. We dealt with first order partial derivatives only. The linearization is legitimate as long as we have small errors and for joint confidence levels below a certain value, usually about 90%. We obtained the general expressions for the parameter errors in the case of fitting a large number, say N, of measurements, for estimating a number of M parameters. The expressions of the parameter errors are functions of the partial derivatives of the measurands with respect to the parameters to be estimated and the standard deviations of the measurements (σ_i).

* Corresponding author. Tel.: +40 21 457 4467x1619; fax +40 21 457 4467.
E-mail address: petre.logofatu@inflpr.ro (P.C. Logofătu).

0263-2241/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.measurement.2008.03.001

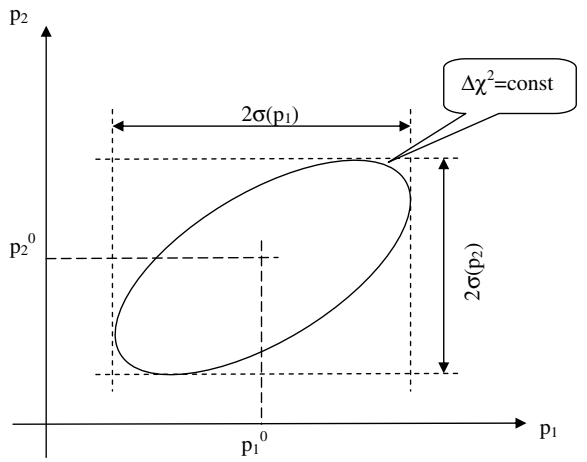

Fig. 1. The ellipsoid of equal probability ($\Delta\chi^2$ isocontour). For two dimensions we have an ellipsis. The errors are the maximum dimensions of the ellipsoid as indicated in the figure.

Let us denote the set of measurands by q_i, $i = 1,\ldots,N$ and the parameters to be estimated by p_j, $j = 1,\ldots,M$. For convenience we will sometimes use the notation

$$\bar{p} \equiv (p_1\ p_2\ \cdots\ p_M)^{\mathrm{T}}, \tag{1}$$

where the superscript "T" means transposition. The measurands q_i are functions of the parameters p_j. Obviously N has to be larger than or equal to M. The fitting procedure requires that we search for the curve that is the best fit of the experimental data. The corresponding parameters p_j^0 are the most probable parameters. That is to say p_j^0 are the parameters for which the function

$$\chi_N^2(\bar{p}) = \sum_{i=1}^{N} \frac{\left(q_i(\bar{p}) - q_i^{\exp}\right)^2}{\sigma_i^2} \tag{2}$$

reaches its minimum, where $q_i^{\exp}$ are the experimentally measured values. Then at $\bar{p} = \bar{p}_0$ (in this entire article the derivations with respect to the parameters p_j are done at $\bar{p} = \bar{p}_0$) we have

$$\frac{\partial\chi_N^2}{\partial p_j} = 0, \quad j = 1,\ldots,M. \tag{3}$$

This is the same as

$$\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left(\frac{\partial q_i}{\partial p_j}\Delta_i^0\right) = 0, \quad j = 1,\ldots,M, \tag{4}$$

where

$$\Delta_i^0 = q_i(\bar{p}_0) - q_i^{\exp}, \tag{5}$$

with the associated probability density of

$$\mathcal{P}\left(\frac{\Delta_i^0}{\sigma_i}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\Delta_i^{0^2}}{2\sigma_i^2}\right) \tag{6}$$

$\Delta_i^0, i = 1,\ldots,N$ is the set of measurement errors associated with the set of measured values $q_i^{\exp}$, and the fitted parameter values $\bar{p}_0$. The error sets like those defined in Eq. (5) form an $\mathfrak{R}^{N-M}$ subspace of the $\mathfrak{R}^N$ error space, because M constraints are imposed upon Δ_i^0 in Eq. (3). Therefore this subspace can be parameterized with N–M parameters. Since $\bar{p}_0$ has a determined, unique value, basically, the N–M degrees of freedom of this parameterization means that $\bar{q}^{\exp}$ can have only values that satisfy Eq. (3). We are not going to do this parameterization, because it would serve no purpose, but it is important to note the possibility of it. This parameterization takes care of N–M dimensions out of the N dimensions of the error space. A complementary parameterization of the remaining M dimensions combined with the first parameterization would be equivalent to a change of coordinates.

The same way as in Eq. (5) we define another set of measurement errors, the set corresponding to the situation when the values of the parameters are not $\bar{p}_0$ but $\bar{p}$, for a any set $\bar{q}^{\exp}$ that satisfies Eq. (3)

$$\Delta_i = q_i(\bar{p}) - q_i^{\exp}, \quad i = 1,\ldots,N. \tag{7}$$

This is a class of error sets with the M degrees of freedom of $\bar{p}$. If q_i would be a linear function of $\bar{p}$, (7) would be precisely a complementary parameterization like the one we were talking about above. For a small departure from $\bar{p}_0, q_i$ have an approximately linear dependence on $\bar{p}$. For a value of $\bar{p}$ far from $\bar{p}_0$ the linearity is not valid anymore, of course, but the corresponding errors have low probability and they can be discarded. The complementarity results from the linearity in $\bar{p}$ of the errors of the type showed in Eq. (7) and the fact that $\bar{p}$ is independent of the parameters of the first parameterization, for the errors of the type shown in Eq. (5). This can be seen from the fact that the first parameterization correspond to a single value of $\bar{p}$, namely $\bar{p}_0$. We will come back below to this issue with a geometric, intuitive illustration. The complementarity of the two parameterizations is a key feature of this contribution that, on the one hand, it gives to the contribution a marked statistical character by allowing it to cover the whole ensemble of measurement error sets, and, on the other hand, makes possible the ensuing derivations by greatly simplifying the algebra. Since the following considerations in which we used the second parameterization are valid for every set $q_i^{\exp}$, then we have taken into account all the possible error measurement sets.

Obviously we have

$$\chi_N^2(\bar{p}) = \sum_{i=1}^{N} \frac{\Delta_i^2}{\sigma_i^2}. \tag{8}$$

Using a linearized Taylor expansion we can express χ_N^2 as

$$\begin{aligned}\chi_N^2(\bar{p}) &= \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left(q_i^0 - q_i^{\exp} + \sum_{j=1}^{M}\frac{\partial q_i}{\partial p_j}\delta p_j\right)^2 \\ &= \sum_{i=1}^{N}\frac{\Delta_i^{0^2}}{\sigma_i^2} + 2\sum_{j=1}^{M}\delta p_j \sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(\frac{\partial q_i}{\partial p_j}\Delta_i^0\right) \\ &\quad + \sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(\sum_{j=1}^{M}\left(\frac{\partial q_i}{\partial p_j}\delta p_j\right)\right)^2 \\ &= \chi_N^2(\bar{p}_0) + \sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(\sum_{j=1}^{M}\left(\frac{\partial q_i}{\partial p_j}\delta p_j\right)\right)^2.\end{aligned} \tag{9}$$

In (9) we used (4) for simplifying the expression. It is more suitable to use a different function instead of χ^2, namely