electronic reprint

Journal of
Synchrotron Radiation
ISSN 0909-0495

Statistical evaluations in fitting problems

K. V. Klementev

Statistical evaluations in fitting problems

K. V. Klementev

Moscow Engineering Physics Institute, 115409 Kashirskoe sh. 31, Moscow, Russia. E-mail: klmn@htsc.mephi.ru

The problem of error analysis is considered taking into account all possible correlations and a prior information about the accessible parameter space. Special attention is paid to the correct determination of the relative weight of experimental data and the *a priori* guess. The applications of statistical χ^2- and F-tests to the fitting problems are also discussed.

Keywords: error analysis; statistical tests.

1. Introduction

Although there exist many textbooks on general data analysis, the problem of statistical evaluations of the errors is under continuous discussion within the EXAFS community. Practically all known programs for EXAFS modelling (IXS, 2000) in some way calculate confidence limits of fitting parameters. However, since there is no standardized technique for that and since the most part of published EXAFS works do not contain any mention of the methods for estimation of the errors of fitting parameters, the accuracy of the EXAFS results remains to be a field for trickery.

Recently, the use of Bayesian analysis for XAFS has been extensively discussed by Krappe & Rossner (2000). Their approach includes not only the explicit dependence of the function minimized on the fitting parameters but also indirect dependence via calculated quantities (amplitudes, phases, mean free path, etc.). In addition, for the direct best fitting without Fourier filtration, they have considered also the truncation error in the EXAFS formula sum. Having included all this, they apply the Bayesian analysis in which the most important problem is to correctly determine the relative weight of experimental data and the *a priori* guess given by the regularization parameter α. They proposed, following Turchin *et al.* (1971), two different prescriptions for that. In this paper we propose an alternative formulation for one of them and show that the other is met only near the singularity point. To do this, we repeat briefly the standard Bayesian arguments with focusing on the problem of α determination.

We also discuss the grounds and usage of the statistical tests which can be and have been misused. The special attention was focused on that where and how one can embellish the results and artificially facilitate the statistical tests to be passed.

2. Errors in determination of fitting parameters

Let for the experimental curve $\mathbf{d}$, defined on the mesh of L nodes with errors ε_i, there exists a model $\mathbf{m}$ that depends on N-dimensional parameter vector $\mathbf{p}$. In EXAFS fitting problems as $\mathbf{d}$ may serve $\chi(k)$, filtered $\chi(k)$, or $\chi(r)$. The problem is to find the parameter vector $\hat{\mathrm{p}}$ that gives the best coincidence of the experimental and model curves. Introduce the figure of merit, the χ^2-statistics:

$$\chi^2 = \frac{N_{\rm ind}}{L}\sum_{i=1}^{L}\frac{(d_i - m_i)^2}{\varepsilon_i^2}, \qquad (1)$$

where $N_{\rm ind}$ is the number of independent experimental points (Stern, 1993). The variate χ^2 follows the χ^2-distribution with $N_{\rm ind} - N$ degrees of freedom.

Let us now derive the expression for the posterior distribution for an arbitrary fitting parameter p_j:

$$P(p_j|\mathbf{d}) = \int d\mathbf{p}_{\neq j} P(\mathbf{p}|\mathbf{d}), \qquad (2)$$

where $P(\mathbf{p}|\mathbf{d})$ is the joint posterior probability density function for all values $\mathbf{p}$, and the integration is done over all $\mathbf{p}$ except for p_j. According to Bayes theorem,

$$P(\mathbf{p}|\mathbf{d}) \propto P(\mathbf{d}|\mathbf{p})P_{\rm prior}(\mathbf{p}), \qquad (3)$$

$P_{\rm prior}(\mathbf{p})$ being the joint prior probability for all $\mathbf{p}$. Assuming that $N_{\rm ind}$ out of L values $\mathbf{d} - \mathbf{m}$ are independent and normally distributed with zero expected values and the standard deviations ε_i, the probability $P(\mathbf{d}|\mathbf{p})$, so-called likelihood function, is given by $P(\mathbf{d}|\mathbf{p}) \propto \exp\left(-\chi^2/2\right)$. The expansion of χ^2 in terms of $\mathbf{p}$ near the minimum, χ_0^2, $(\nabla_p\chi^2 = 0)$ which is reached at $\mathrm{p} = \hat{\mathrm{p}}$ yields:

$$P(\mathbf{d}|\mathbf{p}) \propto \left(-\frac{1}{4}\sum_{k,l=1}^{N} H_{kl}\Delta p_k \Delta p_l\right), \qquad (4)$$

where $\Delta p_k = p_k - \hat{p}_k$, and the Hessian H components (the second derivatives of χ^2) are calculated by the fitting program at $\chi^2 = \chi_0^2$.

Now, we should define the prior probability. Let the parameter p_k is known to lie within the range of the size S_k. Then the prior probability can be expressed as:

$$P_{\rm prior}(\mathbf{p}|\alpha) \propto \alpha^{N/2}\exp\left(-\frac{\alpha}{2}\sum_{kl=1}^{N} A_{kl}(\Delta p_k)^2\right), \qquad (5)$$

where $A_{kl} = \delta_{kl}S_k^{-2}$. The rationale of this prior is that it maximizes the information theory entropy $-\int P_{\rm prior}\ln P_{\rm prior}d\mathbf{p}$ under the constraints $\int P_{\rm prior}d\mathbf{p} = 1$ and $\langle p_k p_l\rangle_{\rm prior} = \delta_{kl}S_k^2$. In other words, this prior introduces minimum information in addition to the approximate knowledge of the sizes S_k. The regularization parameter α specifies the relative weight of the prior probability; at $\alpha = 0$ there is no prior information, at $\alpha \to \infty$ the fitting procedure gives nothing and the posterior distribution coincides with the prior one. In the expression (5) α appears as known value. Really, α should yet be determined. This problem will be considered below.

Finally, for the posterior probability density functions we have:

$$P(p_j|\mathbf{d},\alpha) \propto \int d\mathbf{p}_{\neq j}\alpha^{N/2}\exp\left(-\frac{1}{2}\sum_{k,l=1}^{N} g_{kl}\Delta p_k \Delta p_l\right), \qquad (6)$$

where $g_{kl} = \alpha\delta_{kl}S_k^{-2} + H_{kl}/2$. The matrix H should be generalized, as was done by Krappe & Rossner (2000), to account for the inaccuracies of the calculated scattering amplitudes and phases and truncation errors.

Now, if α would be known, the standard errors of the fitting parameters could be readily obtained:

$$(\delta p_j)^2 \equiv \langle(\Delta p_j)^2\rangle_{\rm post} = \frac{\int(\Delta p_j)^2 P(p_j|\mathbf{d},\alpha)dp_j}{\int P(p_j|\mathbf{d},\alpha)dp_j} = \sum_{i=1}^{N}\frac{e_{ij}^2}{\lambda_i}, \qquad (7)$$

where λ_i and $\mathbf{e}_i$ are the eigenvalues and corresponding eigenvectors of the matrix g.

The only problem that remains to be solved is to determine the regularization parameter α. It should be noticed that on the one

hand, α specifies the relative weight of the prior information, on the other, it makes the matrix g to be positively defined. In general, the matrix H is not positive, especially when N is great and several ill-conditioned directions in the parameter space appear, or when the model and experimental curves differ essentially. Thus, the regularization guarantees that all λ's in Eq. (7) are positive and essentially not zeros.

In the modern Bayesian methods, α is itself determined by Bayesian arguments that maximize the posterior probability of α given the data (Turchin & Nozik, 1969):

$$P(\alpha|\mathbf{d}) = \int d\mathbf{p} P(\alpha, \mathbf{p}|\mathbf{d}) = \int d\mathbf{p} P(\alpha) P(\mathbf{p}|\alpha, \mathbf{d}). \quad (8)$$

Using a prior $P(\alpha) = \alpha^{-\beta}$ (usually a Jeffreys prior with $\beta = 1$ is used (Jeffreys, 1939)), one obtains the posterior distribution:

$$P(\alpha|\mathbf{d}) \propto (\lambda_1 \cdots \lambda_N)^{-1/2} \alpha^{N/2-\beta}. \quad (9)$$

Having found the maximum of this distribution, one obtains the most probable value of α, $\alpha_{\rm mp}$. Then by Eq. (7) one finds the Bayesian errors of fitting parameters. We have found (the proof would take much space here) that at $\alpha = \alpha_{\rm mp}$

$$\langle \chi^2 \rangle_{\rm post} - \chi_0^2 = \sum_{ikl=1}^{N} \frac{H_{kl}}{2} \frac{e_{ik} e_{il}}{\lambda_i} = 2\beta, \quad (10)$$

and this condition appears to be independent of N and ε_i (and any pre-factor in the definition of the χ^2). Eq. (10) can be considered as the equivalent equation for the maximization of $P(\alpha|\mathbf{d})$.

Further, it is easy to verify that at $\alpha = 0$, and if H is positively defined (so called well-posed case) i.e. if the bare likelihood is normalizable then

$$\langle \chi^2 \rangle_{\rm post} - \chi_0^2 = N. \quad (11)$$

The increase of α narrows the posterior distribution, therefore the posterior average of $\chi^2 - \chi_0^2$ decreases and is always less than N (see Fig. 1, solid line). For the ill-posed case, when the matrix g is not positively defined, Krappe & Rossner (2000) proposed the condition $\langle \chi^2 \rangle_{\rm post} = L$ which we can generalize (for the filtered EXAFS also) as $\langle \chi^2 \rangle_{\rm post} = N_{\rm ind}$. We have reservations about the rationality of this condition. (i) This condition follows from the Eq. (11) only when $\chi_0^2 = N_{\rm ind} - N$. In practice, very frequently χ^2 function is not only greater than the median $N_{\rm ind} - N$ but even does not obey the χ^2 distribution law (see the next section). (ii) This condition can be met only near the singularity point, where $\langle \chi^2 \rangle_{\rm post}$ sharply changes from $+\infty$ to $-\infty$, and we wish to pay attention to this latent circumstance. Furthermore, (iii) at that α the matrix g is not yet positively defined (see Fig. 1), and the posterior distribution is divergent.

Returning to the maximization of the posterior probability $P(\alpha|\mathbf{d})$, we point out that if we would choose a uniform prior for α, i.e. with $\beta = 0$, as in Ref. (Krappe & Rossner, 2000), the r.h.s. of Eq. (10) would be zero and the finite solution can be found only for the ill-posed case near the singularity point, as seen from Fig. 1. Thus, both prescriptions proposed in Ref. (Krappe & Rossner, 2000) are not general because for a well-posed case they give only trivial solutions $\alpha = 0$ and $\alpha = \infty$. For general case, the regularization parameter should be found from Eq. (9) or Eq. (10) with nonzero β.

Finally, we discuss briefly the usage of the prior information and give some example results. Very often in the Bayesian applications, the matrix A is considered as the unit matrix (Krappe & Rossner, 2000). Of course, this significantly simplifies the analysis since A and H commute. However, such matrix does not correspond to any feasible information, it just makes the matrix g, with appropriate α, to be positively defined. Contrarily, the prior probability (5) not only regularizes possibly ill-conditioned directions but also expresses the prior knowledge about the accessible parameter space. In our example fitting of the Fourier filtered $\chi(k) \cdot k^2$ with $N = 7$, we have found the errors of the first coordination sphere radius: neglecting all correlations, $\delta^{(a)} r_1 = (2/H_{r_1 r_1})^{1/2} = 7.6 \cdot 10^{-3}$ Å; calculating all correlations and not including prior information (by Eq. (7) with $\alpha = 0$), $\delta^{(b)} r_1 = 3.0 \cdot 10^{-1}$ Å; including the prior information that r_1 lies within ± 0.2Å from the crystallographically determined distance (by Eq. (7) with $\alpha = \alpha_{\rm mp}$) $\delta^{(c)} r_1 = 4.8 \cdot 10^{-3}$ Å. Though, as well-known, these δ's are directly dependent on ε_i, we do not discuss here the problem of ε_i determination and the δ's are given for the comparison between themselves.

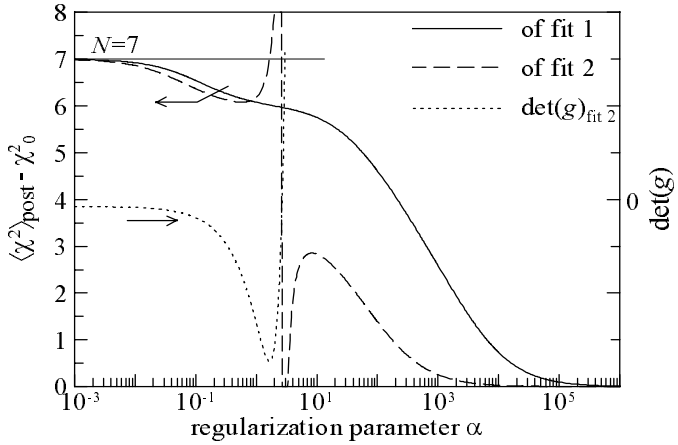

Figure 1
The posterior average of $\chi^2 - \chi_0^2$ as a function of α for two example fittings: solid line for a well-done fitting, dashed line for a bad fitting with initially (at small α) ill-determined matrix g.

3. Statistical tests in fitting problems

3.1. χ^2-test

Introducing the statistical function χ^2, we assumed that it follows the χ^2 distribution with $\nu = N_{\rm ind} - N$ degrees of freedom. However for this would be really so, one should achieve a sufficient fitting quality. This "sufficient quality" could be defined as such that the variate (1) obeys the χ^2 distribution law, that is this variate does not fall within the tail of this distribution. Strictly speaking, the following condition must be met:

$$\chi^2 < (\chi^2_\nu)_\ell, \quad (12)$$

where the critical value $(\chi^2_\nu)_\ell$ for the specified significance level ℓ may be calculated exactly (for even ν) or approximately (for odd ν) using the known formulas (Abramowitz & Stegun, 1964).

Notice, that the choice of the true ε_i here also plays a cardinal role. However, it is important here that one would not use the overestimated values which facilitate to meet the requirement (12). For example, one could obtain the overestimated ε_i, having assumed the Poisson distribution law for the detectors counts when the actual association between the probability of a single count event and the radiation intensity is unknown.

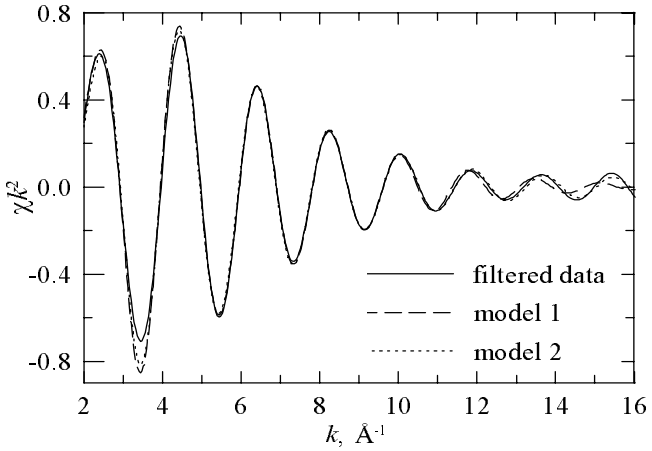

Figure 2
On the choice between two different models on statistical grounds. Cited from Ref. (Menushenkov & Klementev, 2000).

Thus, the exaggerated values ε_i tell about a quality fitting, but give the large errors of fitting parameters. The understated ε_i lead to the would-be small errors, but make difficult to pass the χ^2-test (i.e. to meet the condition (12)). We are aware of many works the authors of which do not describe explicitly the evaluation process for the errors of EXAFS-function extraction and do not report their explicit values. However, by implication it is seen that ε_i were chosen (not calculated) as low as possible to scarcely (with $\ell = 0.9-0.95$) pass the χ^2-test; as a result, very impressive errors of the structural parameters were obtained.

3.2. *F*-test

Let there is a possibility to choose between two physical models depending on different numbers of parameters N_1 and N_2 ($N_2 > N_1$). Which one of them is more statistically important? For instance one wish to decide whether a single coordination sphere is split into two.

Let for the two models the functions χ_1^2 and χ_2^2 obey the χ^2-distribution law with $\nu_1 = N_{\rm ind} - N_1$ and $\nu_2 = N_{\rm ind} - N_2$ degrees of freedom, correspondingly. From the linear regression problem (near the minimum of χ^2, the likelihood function is expressed by (4) and is identical in form to that of the linear regression problem) it is known that the value

$$f = \frac{(\chi_1^2 - \chi_2^2)/(\nu_1 - \nu_2)}{\chi_2^2/\nu_2} \qquad (13)$$

obeys the Fisher's F-distribution law with $(\nu_1 - \nu_2, \nu_2)$ degrees of freedom if exactly $r = \nu_1 - \nu_2$ parameters in the second model are linearly dependent, that is if exist the $r \times N_2$ matrix C of rank r and the vector $\mathbf{c}$ of the dimension r such that $C\mathbf{p} = \mathbf{c}$. In order for the linear restrictions on the second model parameters to be absent, the value f should *not* follow the F-distribution, that is it should be greater than the critical value $(F_{\nu_1-\nu_2,\nu_2})_\ell$ for the specified significance level ℓ: $f > (F_{\nu_1-\nu_2,\nu_2})_\ell$, or

$$\chi_2^2 < \chi_1^2 \left((F_{\nu_1-\nu_2,\nu_2})_\ell \frac{\nu_1 - \nu_2}{\nu_2} + 1 \right)^{-1}. \qquad (14)$$

Notice, that the expression (14) means the absence of exactly r linear restrictions on the second model parameters. Even if (14) is realized, the less number of linear dependencies are possible. If, for instance, the splitting of a single coordination sphere into two does not contradict to the F-test (14), some of the parameters of these two spheres may be dependent, but not all. This justifies the introduction of a new sphere into the model EXAFS function.

Thus, having specified the significance level ℓ, one can answer the question "what decrease of χ^2 must be achieved to increase the number of parameters from N_1 to N_2?" or, inside out, "what is the probability that the model 2 is better than the model 1 at specified (N_1, χ_1^2) and (N_2, χ_2^2)?"

Notice, that since in the definition for f the ratio χ_1^2/χ_2^2 appears, the actual values of ε_i become not important for the F-test (only if they all are taken equal to a single value).

Consider an example of the statistical tests in the fitting problem. In Fig. 2 are shown the experimental curve with $N_{\rm ind} = 11.8$ and two model curves with $N_1 = 4$ and $N_2 = 7$. The underlying physical models were described in Ref. (Menushenkov & Klementev, 2000); here only the number of parameters is of importance. Let us apply the statistical tests. Through the fitting procedure for the model 1 we have: $\nu_1 = 11 - 4 = 7$, $\chi_1^2 = 16.8 > 14.1 = (\chi_7^2)_{0.95}$, for the model 2: $\nu_2 = 11 - 7 = 4$, $\chi_1^2 = 5.3 < 9.5 = (\chi_4^2)_{0.95}$. That is the first model does not pass the χ^2-test. Further, $f = 2.89 = (F_{3,4})_{0.84}$, from where with the probability of 84% we can assert that the model 2 is better than the model 1.

In the EXAFS analysis the F-test has long been in use (Joyner *et al.*, 1987). However, the words substantiating the test are often wrong. The authors of Refs. (Filipponi & Cicco, 1995; Michalowicz *et al.*, 1999) even claimed that the value f from Eq. (13) *must* follow the F-distribution, although then in Ref. (Michalowicz *et al.*, 1999) there appears the correct inequality (14).

4. Conclusion

Though the analysis methods described in this paper are quite standard, they contain several bottlenecks or contradictory points. We have tried to clarify the problem of the regularization parameter determination in the Bayesian approach as well as the grounds for the statistical tests.

This work was supported by RFBR (99-02-17343) and Program "Superconductivity" (99010). I thank the Program Committee for the financial support of my participation in the XAFS-XI conference.

References

Abramowitz, M. & Stegun, I. (eds.) (1964). *Handbook of mathematical functions with formulas, graphs and mathematical tables.* National bureau of standards: Applied mathematical series, 55.

Filipponi, A. & Cicco, A. D. (1995). *Phys. Rev. B,* **52**, 15135–15149.

IXS, (2000). Catalog of XAFS Analysis Programs, `http://ixs.csrri.iit.edu/catalog/XAFS_Programs`.

Jeffreys, H. (1939). *Theory of Probability*. London: Oxford University Press. Later editions: 1948, 1961, 1983.

Joyner, R. W., Martin, K. J. & Meehan, P. (1987). *J. Phys. C: Solid State Phys.* **20**, 4005–4012.

Krappe, H. J. & Rossner, H. H. (2000). *Phys. Rev. B,* **61**, 6596–6610.

Menushenkov, A. P. & Klementev, K. V. (2000). *J. Phys.: Condens. Matter*, **12**, 3767–3786.

Michalowicz, A., Provost, K., Laruelle, S. & Mimouni, A. (1999). *J. Synchrotron Rad.* **6**, 233–235. (Proc. of Int. Conf. XAFS X).

Stern, E. A. (1993). *Phys. Rev. B,* **48**(13), 9825–9827.

Turchin, V. F., Kozlov, V. P. & Malkevich, M. S. (1971). *Sov. Phys. Usp.* **13**, 681–840.

Turchin, V. F. & Nozik, V. Z. (1969). *Izv. Atmospheric and Oceanic Physics,* **5**, 29.

 Received 31 July 2000 • Accepted 25 October 2000