STATE RESEARCH CENTER OF RUSSIA

INSTITUTE FOR HIGH ENERGY PHYSICS

IHEP 95-48

S.I.Alekhin*

ON REDUCING SYSTEMATIC ERRORS IN SOME STATISTICAL ANALYSIS

*alekhin@mx.ihep.su

Protvino 1995

UDK 519.254 M–24

Abstract

Alekhin S.I. On reducing systematic errors in some statistical analysis: IHEP Preprint 95-48. – Protvino, 1995. – p. 4, refs.: 4.

We analyse statistical properties of the simplest χ^2 estimator being applied to the analysis of the data correlated due to common systematic uncertainties within the Bayesian approach. Analytical formula for the systematic errors and the bias of the parameter estimator are presented. Stressing that this estimator is not efficient we show that the systematic errors of the fitted parameters can be decreased by using in this task estimators based on the likelihood function whereas the values of the fitted parameters are shifted to the true value. The described effect probably can help resolve some contradictions in the particle phenomenology.

Аннотация

Алехин С.И. Подавление систематических ошибок в некоторых статистических обработках: Препринт ИФВЭ 95-48. – Протвино, 1995. – 4 с., библиогр.: 4.

Мы анализируем статистические свойства простейшей χ^2 оценки при анализе данных с коррелированными систематическими ошибками на основе байесовского подхода. Представлены аналитические формулы для систематических ошибок параметров и смещения оценки. Обращая внимания на то, что данная оценка не эффективна, мы показываем, что систематические ошибки фитируемых параметров могут быть уменьшены при использовании оценок, основанных на функции максимального правдоподобия; при этом значения параметров смещаются в сторону истинного значения. Описанный эффект может помочь объяснению различных противоречий в феноменологии.

Modern particle physics development is often based on the analysis of precise experimental data. One of the well known problem of such analysis is the account of systematic uncertainties of the data which often are comparable or even larger than the statistical ones [1]. It is customary that systematic errors are ignored at the first stage of the analysis and at the second stage data are shifted by the value of a systematic uncertainty and the analysis is repeated to evaluate systematic errors for the parameters of the applied theoretical model (see, for example [2]). At the final stage individual systematic errors can be combined in quadrature. The goal of this letter is to examine basic statistical properties of the parameters estimator obtained using this method within the Bayesian approach.

If data are explicitly described by a theoretical model and in the presence of K sources of systematic errors experimental data can be presented as

$$y_i = f_i + \mu_i \sigma_i + \lambda_k s_i^k, \tag{1}$$

where $f_i = f_i(\theta^0)$ is the value predicted by the theoretical model with parameter θ^0, μ_i and λ_k are independent random variables, σ_i and s_i^k - statistic and systematic errors from the k-th source for i-th measurement, $i = 1 \cdots N$, $k = 1 \cdots K$, N is the total number of points in the data set. If data come from the data sample with a large number of events, μ is normally distributed, as to λ, the only assumption is that they have zero average and unity dispersions. In accordance with the approach analysed here one finds the estimator of the parameter $\hat{\theta}$ by minimization of χ^2 functional

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(f_i(\theta) - y_i)^2}{\sigma_i^2}.$$

To obtain the dispersion of $\hat{\theta}$ we will follow the method used in [3]. Following their notations we introduce the quantities

$$\xi(\theta) = \frac{\partial \chi^2}{\partial \theta}, \qquad X = \xi(\theta^0),$$

$$a = - < \frac{\partial \xi(\theta^0)}{\partial \theta} >,$$

$$b = < \frac{\partial^2 \xi(\theta^0)}{\partial \theta^2} >, \qquad Y = \frac{\partial \xi(\theta^0)}{\partial \theta} + a,$$

where $< >$ means averaging over the repeated data samples. In these notations

$$\hat{\theta} - \theta^0 = \frac{X}{a} + \frac{XY}{a^2} + \frac{bX^2}{2a^3} + \cdots, \tag{2}$$

and the rejected part of the expansion contains terms with the higher powers of $1/a$. In this approximation and neglecting σ_i fluctuations dispersion and bias of $\hat{\theta}$ can be expressed as

$$D(\hat{\theta}) = \frac{< X^2 >}{a^2},$$

$$B(\hat{\theta}) = \frac{< XY >}{a^2} + \frac{< bX^2 >}{2a^3}.$$

If data points are uncorrelated averaging of X^2 and XY leads to the cancellation of most terms in the double sum and one can obtain

$$< X^2 >= \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial f_i(\theta^0)}{\partial \theta} \frac{\partial f_i(\theta^0)}{\partial \theta} = -a,$$

$$< XY >= \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial f_i(\theta^0)}{\partial \theta} \frac{\partial^2 f_i(\theta^0)}{\partial \theta^2} = \frac{b}{3}.$$

With the account of the data correlations the expression becomes more complicated

$$< X^2 >= \sum_{i,j=1}^{N} \frac{C_{ij}}{\sigma_i^2 \sigma_j^2} \frac{\partial f_i(\theta^0)}{\partial \theta} \frac{\partial f_j(\theta^0)}{\partial \theta} = -a + \Big[\sum_{k=1}^{K} \sum_{i=1}^{N} s_i^k \frac{\partial f_i(\theta^0)}{\partial \theta} \Big]^2,$$

$$< XY >= \sum_{i,j=1}^{N} \frac{C_{ij}}{\sigma_i^2 \sigma_j^2} \frac{\partial f_i(\theta^0)}{\partial \theta} \frac{\partial^2 f_j(\theta^0)}{\partial \theta^2} =$$

$$= \frac{b}{3} + \Big[\sum_{k=1}^{K} \sum_{i=1}^{N} s_i^k \frac{\partial f_i(\theta^0)}{\partial \theta} \Big] \Big[\sum_{k=1}^{K} \sum_{i=1}^{N} s_i^k \frac{\partial^2 f_i(\theta^0)}{\partial^2 \theta} \Big],$$

where C_{ij} is the correlation matrix

$$C_{ij} = \sum_{k=1}^{K} s_i^k s_j^k + \delta_{ij} \sigma_i \sigma_j$$

and δ_{ij} is Kronecker symbol. At first we should note that the last formula gives an analytical result for the dispersion and bias of $\hat{\theta}$ due to systematical errors as they are treated in the considered approach and can be used to save computer time by omitting fit repetitions. The second note is that the nominator of the expression for the dispersion is $\sim N^2$ whereas without correlations it is $\sim N$. This unpleasant property is also reproduced in higher terms of expansion (2) and leads to increase of the estimator bias

and to worse statistical convergence of this estimator with increasing the data amount. The reason of such worsening is obvious and is connected with the fact that with the account of correlations minimization of the simple χ^2 functional becomes nonequivalent to the maximizing likelihood function which gives the efficient estimator and hence the dispersion of $\hat{\theta}$ becomes larger than Cramer-Rao limit. For the given distribution of λ_k one can construct this likelihood function. Say if one supposes that they are normally distributed, the most optimal estimator is provided by the minimization of the functional

$$\chi^2(\theta) = \sum_{i,j=1}^{N} q_i E_{ij} q_j, \tag{3}$$

where $q_i = f_i(\theta) - y_i$, and E_{ij} is inverted correlation matrix. Dispersion of $\hat{\theta}$, i.e. systematic errors of the fitted parameter in this case is

$$D(\hat{\theta}) = \left[\sum_{i,j=1}^{N} \frac{\partial f_i(\theta^0)}{\partial \theta} E_{ij} \frac{\partial f_j(\theta^0)}{\partial \theta} \right]^{-1}. \tag{4}$$

As far C_{ij} is positive definite one can construct real matrix $\sqrt{C}_{ij}$ giving its square root and using triangle inequality we obtain regardless the shape of λ distribution

$$\left[\sum_{i,j=1}^{N} \frac{C_{ij}}{\sigma_i^2 \sigma_j^2} \frac{\partial f_i(\theta^0)}{\partial \theta} \frac{\partial f_j(\theta^0)}{\partial \theta} \right] \left[\sum_{i,j=1}^{N} \frac{\partial f_i(\theta^0)}{\partial \theta} E_{ij} \frac{\partial f_j(\theta^0)}{\partial \theta} \right] =$$

$$= \left[\sum_{i=1}^{N} \left(\sum_{j=1}^{N} \frac{\sqrt{C}_{ij}}{\sigma_j^2} \frac{\partial f_j(\theta^0)}{\partial \theta} \right) \left(\sum_{j=1}^{N} \frac{\sqrt{C}_{ij}}{\sigma_j^2} \frac{\partial f_j(\theta^0)}{\partial \theta} \right) \right] \times$$

$$\times \left[\sum_{i=1}^{N} \left(\sum_{j=1}^{N} \sqrt{E}_{ij} \frac{\partial f_j(\theta^0)}{\partial \theta} \right) \left(\sum_{j=1}^{N} \sqrt{E}_{ij} \frac{\partial f_j(\theta^0)}{\partial \theta} \right) \right] \geq$$

$$\geq \left[\sum_{i,j,l=1}^{N} \frac{\sqrt{C}_{ij}}{\sigma_j^2} \frac{\partial f_j(\theta^0)}{\partial \theta} \sqrt{E}_{il} \frac{\partial f_l(\theta^0)}{\partial \theta} \right]^2 = a^2.$$

This means that even if the systematic errors are not Gaussian distributed the full treatment of the correlations can give the estimator with smaller dispersion, though it can be not efficient in this case. From the above one can see that the difference is the greater the larger N is and the effect of reducing the systematics can be more significant for global data analysis. Generalization on multi-parametric case is obvious. In this case $1/a$ is replaced by the error matrix for the parameters and in the case of large correlations of the parameters the effect of systematic errors reduction can be more pronounced.

One can argue that the Gaussian distribution of systematic errors is not a very strong assumption (see, for example, discussion in [4]). If the systematic error arises from the poor knowledge of some parameter of the experimental apparatus (geometrical dimensions, counter efficiency, etc.) it is rather natural to suppose the Gaussian distribution

for the scale of this error which is obtained as the propagation of the error for this parameter. In the cases when the systematic error is evaluated as the error in correction which is calculated using Monte-Carlo generation this approximation seems to be almost correct. If K is large, y_i tends to obey the Gaussian distribution for any distribution of λ in accordance with the central limit theorem of statistics and this approximation is better if s_i^k have comparable values. Thus ansatz (3) with corresponding formula (4) in many cases can not only improve dispersion of estimator, but provide its efficiency.

If the effect of the systematics reduction is large, the data reanalysis with full treatment of correlations can also lead to a significant shift of the parameters. This shift is of the order of value of systematic error evaluated with the simplest estimator and since the dispersion of the new estimator is reduced, the new value of the parameter comes closer to the true value. The effect is more pronounced if the simplest estimator suffer from the bias. One can hope that the described statistical approach can help resolve some contradictions in the particle phenomenology.

The author is grateful to A.S.Nikolaev for reading the manuscript and valuable comments.

References

[1] D'Agostini G.-DESY-93-175, 1993; Schmelling Michael.-CERN/PPE-94-185, Nov. 1994.

[2] Benvenuti A.C. et al. //Phys. Lett. 1989. V.223B. P.490; Benvenuti A.C. et al. //Phys. Lett. 1989. V.237B. P.592; Arneodo M. et al. // Phys. Lett. 1993. V.308B. P.222; Lancaster Mark.-DESY-94-204, P.15, 1994.

[3] Eadie W.T., Drijard D., James F.E., Roos M., Sadoulet B. Statistical Methods in Experimental Physics, North Holland, 1971.

[4] Soper Davison E., Collins John C.-CTEQ NOTE 94/01, Nov. 1994.

Received March 3, 1995

С.И.Алехин
Подавление систематических ошибок в некоторых статистических обработках .

Оригинал-макет подготовлен с помощью системы LaTeX.
Редактор Е.Н.Горина. Технический редактор Н.В.Орлова.

Подписано к печати 20.04.1995 г. Формат 60 × 84/8. Офсетная печать.
Печ.л. 0,5. Уч.-изд.л. 0,4. Тираж 240. Заказ 275. Индекс 3649.
ЛР №020498 06.04.1992 г.

ГНЦ РФ Институт физики высоких энергий
142284, Протвино Московской обл.

Индекс 3649

П Р Е П Р И Н Т 95-48, И Ф В Э, 1995