Vol 13 No 4, April 2004
1009-1963/2004/13(04)/0428-06

Chinese Physics

© 2004 Chin. Phys. Soc.
and IOP Publishing Ltd

A new method to avoid PPP in data evaluation*

Chen Zhen-Peng(陈振鹏)† and Sun Ye-Ying(孙业英)

Tsinghua University, Beijing 100084, *China*

(Received 30 May 2003; revised manuscript received 13 October 2003)

The Peelle Pertinent Puzzle (PPP) has been studied systematically. The results obtained by using several proposed methods to deal with a sample of ^{6}Li(n,t) are compared with each other. The cause of PPP occurrence has been found. A new method to avoid PPP is introduced, whose main idea is 'to selectively make suitable correction for systematic error according to the relative extent of discrepancy'. The method is able to produce the best result and should be used to avoid the PPP at first.

Keywords: general least squares, PPP, data evaluation
PACC: 0250, 2980

1. Introduction

The most widely used method of parameter estimation in statistics is that of general least squares (GLS). But in some cases, the estimate of GLS has an evident bias below the majority of experimental data, which has been called Peelle's Pertinent Puzzle (PPP) [1] in the nuclear community. Recently, in the evaluation of the standard cross section of ^{6}Li(n,t) [2] the PPP occurred (referred to as Gmar-VP in Fig. 1). Some approaches [1–6] have been proposed to solve this problem, but viewpoints on this issue diverge significantly, and no consensus exists at present on how to deal with this specific problem.

We have studied the PPP problem systematically by using all the methods proposed earlier and also a new method proposed in this paper to process the data of the ^{6}Li(n,t) reaction used in Ref.[2]. In this paper, a comparison of the results of all these methods is made, the cause for the occurrence of PPP is studied, and a new method is introduced to avoid PPP.

2. GLS-GMA formula and code C-GLS

Suppose the nonlinear regression model is

$$y = f(\eta, \theta), \tag{1}$$

where y refers to the vector of posterior, η refers to the observable vector, and θ refers to the parameter vector. For many applications it is possible to stipulate that a linear relationship exists between η and θ,

$$y - y_a = G(y, y_a). \tag{2}$$

When θ is just an evaluated vector of η, and y_1, $y_2, \ldots, y_n$ are independent of each other, the general least squares Gauss–Markov–Aitken (GLS-GMA) formalism gives [7]

$$y - y_a = (G^+V^{-1}G)^{-1}G^+V^{-1}(\eta - y_a), \tag{3}$$

$$\Delta M = M' - M = (G^+V^{-1}G)^{-1}, \tag{4}$$

where y_a is a prior vector. G is a matrix of sensitive coefficients, whose elements are either one or zero. M is a covariance matrix of prior data (at first M is a zero matrix). M' is a covariance matrix of posterior evaluation. V is a covariance matrix of experimental data. The superscripts $(+)$ and (-1) mean transpose and inverse operations respectively.

A new code C-GLS was developed based on formulae (3) and (4) to study PPP, and the calculated result of C-GLS is identical to that of the code GMA which has been used widely in the international nuclear data community for many years.

According to the Gauss–Markov theorem, the GLS condition is given by

$$\chi^2 = (\eta - y)^+V^{-1}(\eta - y) = \text{minimum}, \tag{5}$$

*Project supported by the International Atomic Energy Agency (Grant No CPR-12022/R0) and the National Nuclear Data Program (Grant No y416010502).

†Corresponding author. 62782163(o) 62786552(h) E-mail: zhpchen@mail.tsinghua.edu.cn

from which the estimate of θ can be obtained. But, if the statistics of a sample is not good enough, or it is seriously discrepant, the GLS can produce a biased estimate. In this case, $\chi^2=(\eta-y)^+V\ (\eta-y)$ is less meaningful. Let U be a diagonal matrix formed with the diagonal elements of V; the quantity

$$\text{Chi2} = (\eta-y)^+U(\eta-y)/N \tag{6}$$

is more meaningful for measuring the mean square error of the estimate. N in Eq.(6) is the number of data to be fitted. So, the Chi2 being minimum is chosen as a criterion for searching for the best estimate.

3. The fundamental cause for PPP occurrence in non-model fitting

Usually, if an estimate of GLS has an evident bias below the majority of experimental data, the fit model should be changed. For example, in Ref.[8], when the fitting model is changed by using the Fermi distribution with double parameters instead of a single one, the result becomes very good. But, the GLS-GMA fitting is a kind of non-model fit; the parameters to be estimated are just the estimates of experimental data. So the cause of PPP occurrence must come from the structure of data to be fitted.

It was shown in Ref.[5] that PPP may occur when the covariance matrix of experimental data contains such non-diagonal element(s) V_{ij} which are larger than the diagonal element(s) V_{kk}, that is

$$V_{ij} > V_{kk} \text{ for any } i,j,k \text{ with } i \neq j. \tag{7}$$

But, in the five sets of ^{6}Li (n,t) data used in Ref.[2] (called 'the data sets' hereafter), for each set of data, the non-diagonal elements of covariance matrix are all smaller than the diagonal elements. A primary test shows that the PPP occurrence for the data sets is mainly due to the existence of some data points which are highly discrepant from the others. In the data sets, all errors are given in percentages, so that for a datum value higher than its expected one (called 'higher data' hereafter), it has a larger absolute error and a larger associate covariance, and consequently has a relative smaller weight. In contrast, for a datum lower than its expected value (called 'lower data' hereafter), it has a relative larger weight. When both the errors and discrepancy of a data set are sufficiently large, the PPP may occur.

So any method which is able to remove the phenomena mentioned above will have the effect of avoiding PPP. The discrepancy extent of the experimental data is a kind of experimental information; it should be paid attention to in GLS fitting. Thus we propose a new method to avoid PPP, i.e. 'to selectively make suitable correction for systematic error of a datum, which involves higher discrepancy according to their relative extent of discrepancy'.

Define a factor f_i to describe the extent of discrepancy. First let

$$D_i = (y_{\text{exp},i} - y_{\text{eva},i})/y_{\text{eva},i}, \tag{8}$$

where $y_{\text{exp},i}$ refers to the experimental data, and $y_{\text{eva},i}$ refers to the corresponding evaluated value. Then let

$$f_i = D_i/R_i, \tag{9}$$

where R_i is the relative standard error of the ith datum point. The f_i describes the extent of discrepancy for the ith data point, which can be used as a measurement to make selected modification. For the data sets of ^{6}Li(n,t), the distribution of f_i is shown in Table 1.

Table 1. Distribution of f_i.

Range of f_i	$f_i \le -4$	$-4 < f_i \le -3$	$-3 < f_i \le -2$	$-2 < f_i \le -1$	$-1 < f_i \le 0$	$0 < f_i \le 1$	$1 < f_i \le 2$	$2 < f_i \le 3$	$3 < f_i \le 4$	$4 < f_i$
No. of data	0	3	13	25	37	27	20	15	6	8
Percent/%	0.0	2.0	8.5	16.3	24.2	17.6	13.1	9.8	3.9	5.3

It can be seen from Table 1 that the distribution is far away from the normal distribution; sometimes this kind of sample is called a 'diseased sample'.

In the code C-GLS, the first formula for modification of experimental error is

$$e_i = E_i(1 + C(y_{\text{exp},i} - y_{\text{eva},i})/y_{\text{eva},i}), \tag{10}$$

where E_i is the original error, e_i is the modified one, C is a factor and its value will be assigned by iteration procedure. The second formula is

$$e_i = E_i(1 + C). \tag{11}$$

4. Comparison of the results obtained by different methods

A comparison of the results obtained by all the proposed methods is shown in Table 2. The first column headed 'Title' refers to the name of a calculated result. The second column headed 'Fac-C' refers to the factor C involved in formulae (10) and (11). The third column headed 'Er-type' refers to the type of error corrected in calculation, in which Y means systematic error, S statistic error, YS both systematic and statistic errors, NO means no error being corrected. The fourth column headed 'Thr-F_i' refers to the threshold assigned to F_i ($F_i = |f_i|$ hereafter) to selectively make correction for the relative error, in which 0 means all data are involved; 1, 2 and 0–1 mean the data with $F_i > 1$, $F_i > 2$ and $F_i < 1$ are involved respectively. The fifth column headed 'Chi2' refers to the quantity calculated using formula (6). The sixth column headed 'Rel-CS' refers to the averaged value of all calculated ${}^6\mathrm{Li}(n,t)$. The seventh column headed 'Rel-Cov' refers to the averaged value for all elements of calculated covariance matrix. In the eighth and ninth columns headed 'n–Er (%)', '$0 < F_i \leq 1$' and '$F_i > 1$', in which n is the number of data involved with $0 < F_i \leq 1$ and $F_i > 1$, Er (%) is the change in term of percentage error.

In the first line, labelled 'Wei-Ave', the calculated results are obtained by neglecting the non-diagonal elements of the covariance matrix of fitted data. In fact, the estimates are weighted averaged values. The number 3.703 is found to be the minimum one in all calculated Chi2, so it is taken as the first standard (S_c=3.703) for comparison.

In the second line, labelled 'Gmar-VP', the calculated results are obtained by using all the original experimental errors. In this case, the PPP occurs and the quantity Chi2 (11.11) is very large, but the quantity Rel-Cov is found to be the minimum one in all calculated covariance matrix, so it is taken as the second standard (S_r=0.907) for comparison.

In the third to ninth lines are given the results of calculations, in which the errors of the data are corrected using formula (11).

Table 2. Comparison for summary results of calculated values.

	Title	Fac-C	Er-ty	Thr-F_i	Chi2	Rel-CS	Rel-Cov $0 < F_i \leq 1$	n–Er(%) $F_i > 1$	n–Er(%)
1	Wei-Ave	NO	NO	NO	3.703	1.000	0.140	77–0.00	76–0.00
2	Gmar-VP	NO	NO	NO	11.11	0.907	1.000	77–0.00	76–0.00
3	Md-F-1	7.82%	YS	0	3.787	0.998	1.042	76–7.82	78–7.82
4	Md-F-2	7.82%	Y	0	3.801	0.992	1.037	76–7.82	78–7.82
5	Md-F-3	7.82%	S	0	13.13	0.895	0.963	76–7.82	78–7.82
6	Md-F-4	8.36%	Y	0	3.767	0.999	1.054	79–8.36	75–8.36
7	Md-F-5	8.36%	Y	1	3.870	0.988	1.071	62–0.00	92–8.36
8	Md-F-6	8.36%	Y	0–1	11.42	0.905	0.990	25–8.36	129–0.00
9	Md-F-7	8.69%	Y	1	3.812	0.991	1.072	60–0.00	94–8.69
10	Md-E-1	0.8200	YS	0	3.843	0.998	1.080	61–1.92	93–7.34
11	Md-E-2	0.8200	YS	1	3.850	0.996	1.074	61–0.00	93–7.34
12	Md-E-3	0.8382	YS	1	3.848	0.997	1.075	61–0.00	93–7.58
13	Md-E-4	0.8733	Y	0	3.761	0.999	1.087	55–1.76	99–7.25
14	Md-E-5	0.8733	Y	1	3.768	0.997	1.077	55–0.00	99–7.25
15	Md-E-6	0.9033	Y	1	3.761	0.999	1.085	55–0.00	99–7.43
16	Md-E-7	1.0940	Y	2	3.758	1.000	1.055	42–13.2	(for $F_i > 2$)
17	Chiba	NO	YS	0	4.027	1.013	1.115	65–2.24	89–8.90
18	Zhao.	NO	YS	0	4.017	1.012	1.098	77–1.44	76–8.66
19	Log-Tro	NO	NO	NO	3.940	1.010	1.113	NO	NO
20	Cox–Box	NO	NO	NO	4.220	1.019	1.132	NO	NO
21	RAC-Ch	NO	NO	NO	3.379	1.016	0.971	NO	NO

In the 'Md-F-1' line, both statistical and systematic errors have been corrected by 7.82 % for all data; the result seems good, the cross section is about 0.2 % lower than the first standard S_c, and the covariance is about 4.2 % higher than the second standard S_r. In the 'Md-F-2' line, only systematic error has been corrected by 7.82% for all data, the result seems good too. In the 'Md-F-3' line, only statistical error has been corrected by 7.82 % for all data, the cross section is about 10.5% lower, the result appears very bad, and the PPP occurs. These three calculations show that the correction for systematic error plays the most important role.

In the 'Md-F-4' line, only systematic error has been corrected by 8.36% for all data; the result is improved relative to 'Md-F-2'. In the 'Md-F-5' line, only systematic error has been corrected by 8.36% for the 92 data with $F_i > 1$; the result seems good too. In the 'Md-F-6' line, only systematic error has been corrected by 8.36% for the 25 data with $F_i < 1$; the cross section is about 9.5% lower, the result seems very bad, and the PPP occurs. These three calculations show that the correction for the error of such data with higher discrepancy play the most important role.

In the line 'Md-F-7', only systematic error has been corrected by 8.69% for the 94 data with $F_i > 1$, and the result seems good too.

In the calculated results mentioned above, the best or better ones are the 'Md-F-4' and 'Md-F-7'. It should be noted that the corrected error in these two calculations only involves systematic error.

In the 10th to 16th lines are given the results of calculations in which the errors of the data have been corrected according to the relative extent of discrepancy, and calculated using formula (10).

In the 'Md-E-1' line, both statistical and systematic errors have been corrected with c=0.820 for all the data. For the 61 data with $F_i < 1.0$, accordingly, the error has been corrected by 1.92% on average. For the 93 data with $F_i > 1.0$, the error has been corrected by 7.34% on average. For the total 154 data, the error has been corrected by 4.93% on average, which is lower than the corresponding value (8.72%) of 'Md-F-1'. The result seems good, cross section is only about 0.2% lower, Rel-Cov is about 8.0% higher.

In the 'Md-E-2' line, both the statistical and systematic errors have been corrected with c=0.820 for the 93 data with $F_i > 1$, and accordingly, the error is corrected by 7.34% on average. The result seems good. In the 'Md-E-3' line, both statistic and systematic errors have been corrected with c=0.8382 for the 93 data with $F_i > 1$, and accordingly, the error is corrected by 7.58% on average. The result seems good too. These two calculations show again that the correction for the error of the data with higher discrepancy plays the most important role.

In the 'Md-E-4' line, only systematic error has been corrected with c=0.8733 for all data. In the 'Md-E-5' line, only systematic error has been corrected with c=0.8733 for the 99 data with $F_i > 1$. In the 'Md-E-6' line, only systematic errors has been corrected with c=0.9033 for the 99 data with $F_i > 1$. In the 'Md-E-7' line, only systematic error has been corrected with c=1.094 for the 42 data with $F_i > 2$, and accordingly, the error is corrected by 13.2% on average. All the results mentioned above are clearly improved. The result of 'Md-E-7' may be the best one in all these seven calculations, even though only the systematic error of the 42 data with $F_i > 2$ is corrected. These calculations show again that the correction for systematic error plays the most important role.

In the 17th to 20th lines are given the results calculated using the methods proposed before this work. All the results show that these methods produce 'over-correction' or 'over-estimate' to some extent.

In the 'Chiba' line is the result by using the Chiba method,[4] in which the absolute errors of data have been calculated in terms of given percentage errors and the evaluated values by iteration; the error is corrected by 6.05% on average. The cross section is about 1.3% higher; Rel-Cov is about 11.5% higher. In the 'Zhao' line is the result by using Zhao method,[3] in which the absolute errors of data have been calculated in terms of given percentage errors and the weighted averaged values by iteration; the errors was corrected by 5.83% on average. The cross section is about 1.2% higher; Rel-Cov is about 9.8% higher. In the 'Log-Tro' line is the result of logarithmic transformation,[5] in which the cross section is about 1.0% higher; Rel-Cov is about 11.3% higher. In the 'Cox–Box' line is the result of Cox–Box transformation,[9] in which the cross section is about 1.9% higher; Rel-Cov is about 13.2% higher.

In the 'RAC-Ch' line is the result calculated [10] by means of R-matrix analysis code RAC, in which only the five sets of ^{6}Li(n,t) data was involved. The Rel-Cov is 0.971.

In Fig.1 are shown the five important results calculated by using different methods for the five sets of ^{6}Li(n,t) data. In regard to the result of 'Cox–Box',

the estimate of cross section is about 1.9% higher, and the averaged value for all elements of calculated covariance matrix is about 13.2% higher.

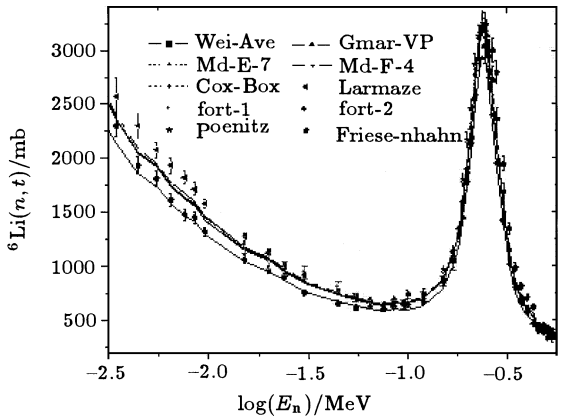


Fig.1. The fit results with five different methods for the five sets of ^{6}Li(n,t) data.

5. Discussion and conclusion

The method to avoid PPP is mainly to find an effective and rational approach to deal with a sample in which the data are seriously discrepant with each other. Usually, the procedure includes:

1) Quantify the degree of confidence on the consistency or inconsistency of the data;

2) Correct or reject possible 'wild' data, if there is a specific reason;

3) Normalize a certain data set, if there is a specific reason;

4) Selectively enhance some of errors of the data, if there is a specific reason.

In the five sets of ^{6}Li(n,t) data, the Friesenhahn data are lower than others systematically for the energy region from 0.0025MeV to 0.1MeV, but are higher than others systematically in a higher energy region. There is about 0.02MeV difference for the positions of the peak $(-5/2)$. There are many very large discrepancies (e.g. $F_i > 2$) in these data sets. Also, no specific reason has been found to consider some data as 'wild' or to normalize them.

Practical calculations show that the PPP occurrence is mainly due to the existence of much higher discrepancy in some data points. So it is reasonable to correct the errors of the data with higher discrepancy. The extent of correction can be controlled by setting a threshold for F_i and by adjusting the factor C to obtain a minimum of Chi2.

The methods of Zhao and Chiba can be in fact thought of as making error correction for all the data. The disadvantage of these two methods is that they are based more or less on subjective interpretation of the fractional uncertainties along with the measured data and have a tendency to produce an 'over-corrected' result.

If a sample has a good normal distribution, its estimate of GLS must be good. The methods of logarithmic and Cox–Box transformation can effectively eliminate the non-normality of a 'diseased sample', so they are effective to overcome PPP for the 'diseased sample'. The problem of these methods is that all the original experimental information is changed to some extent, which could bring out additional bias. For example, for the five sets of ^{6}Li(n,t) data, if the Friesenhahn data are excluded, the first four sets of data will form a sample with very good normal distribution. For this new sample, its estimates of GLS by using original data directly (GMA-Ori) and by using logarithmic transformation (GMA-Log) are listed in Table 3. It can be seen that the logarithmic transformation makes the averaged cross section (Ave-CS) increase by 1.6%, while the averaged covariance (Ave-Cov) increases by 0.9%, and the averaged variance (Ave-Var) increases by 5.1%. The Cox–Box transformation produces a very similar result. For a standard nuclear cross section, this extent of difference is not acceptable.

Table 3. Comparison of the results by using original data and logarithmic transformation.

	Chi2	Ave-CS	Ave-Cov	Ave-Var
GMA-Ori	1.1510	1306.25	235.12	2465.45
GMA-Log	1.0709	1326.61	237.23	2589.83
Log/Ori	0.937	1.016	1.009	1.051

A very important idea [7] is that the validity of Gauss–Markov theorem relies on the existence of mean values and covariance matrices for the *a priori* parameters and the observed data, not on the normalcy of the probability distribution. Thus, for samples with non-normality to some extent, logarithmic and Cox–Box transformation are not the best ways to avoid PPP, because the original samples are changed seriously after transformation, which leads to a tendency to produce an 'over-corrected' result.

Our conclusion is that all the proposed methods are effective to avoid PPP, but Zhao's method , Chiba's method, the logarithmic and Cox–Box transformation would lead to a tendency to produce 'over-

corrected' results. The method 'to selectively make suitable correction for systematic errors according to the relative extent of discrepancy ' has a possibility to produce the best result; it should be used to avoid the PPP at first.

Acknowledgements

The authors would like to thank Dr Liu Tingjin, Dr Zhao Zhixiang of China Institute of Atomic Energy, and Dr V Pronyaev of IAEA for their useful discussions and other help.

References

[1] Chiba S and Smith D L 1991 ANL/NDM-121
[2] Pronyaev V G 2003 IAEA INDC (NDS)- 438 p159
[3] Zhao Z X and Perey F G 1992 ORNL/TM-12106
[4] Chiba S and Smith D L 1994 *J. Nucl. Sci. Technol.* **31** 770
[5] Box G E P and Cox D R 1964 *J. R. Statist. Soc. Series* B **26** 211
[6] Pronyaev V G 2003 IAEA INDC (NDS)- 438 p184
[7] Smith D L 1991 *Probability, Statistics, and Data Uncertainties in Nuclear Science and Technology* (American Nuclear Society) p205
[8] Li J X Guo Z Y and Xiao G Q 2003 *Acta Phys. Sin.* **52** 58 (in Chinese)
[9] Oh S Y 2003 IAEA, INDC (NDS)-438 146
[10] Chen Z P and Sun Y Y 2003 IAEA INDC (NDS)- 438 65