Journal of the Korean Nuclear Society
Volume 28, Number 1, February 1996

A Regression Program COVAFIT Accounting for Variance-Covariances in Experimental Nuclear Data

Soo Youl Oh and Jonghwa Chang
Korea Atomic Energy Research Institute
(Received September 13, 1995)

실험 핵자료의 분산-공분산을 고려한 회귀분석 프로그램 COVAFIT

오수열 · 장종화
한국원자력연구소
(1995. 9. 13 접수)

Abstract

A computer program COVAFIT has been developed and applied to the evaluation of experimental cross sections for MeV energy incident particles. The program utilizes weighted least-square linear regression method with high-order polynomials derived in this study. Meeting the growing demand for the treatment of covariances in nuclear data, it deals with the variance and covariance data provided along with experimental cross sections and yields those for the evaluated ones. The evaluated results on two sets of neutron total cross section of oxygen and three sets of proton cross section for C^{11} production reactions confirm the methodology formulated in and the applicability of the program.

요 약

단면적 평가를 위한 회귀분석 프로그램 COVAFIT를 개발하고 이를 실험 결과로서 보고되는 단면적을 평가하는데 적용하였다. 가중 최소제곱 선형 회귀 방법을 적용하였는데 이 때 새로 유도한 고차 다항식을 사용하였다. 점증하는 핵자료의 공분산에 대한 요구에 부응하여, 이 프로그램은 단면적 실험치와 함께 제공되는 분산-공분산 자료를 평가에 반영할 수 있다는 특징을 가지고 있다. MeV 에너지 영역에서 천연 산소 및 O^{16}의 중성자 전단면적과 세 가지 C^{11} 생산 반응의 양성자 반응 단면적 평가를 통하여 사용 방법론의 적절성과 COVAFIT의 유용성을 확인하였다.

1. Introduction

The nuclear data evaluation in a broad sense includes activities such as the evaluation of point-wise experimental or theoretical data and the processing of the evaluated data to a suitable form for applications. In a narrow sense, this refers to the former activity, and the resulting product is the so-called evaluated nuclear data library such as ENDF of U.S.A, JENDL of Japan, and JEF of Europe. The term 'evaluation' means to generate a single representative value for a fixed condition and, in addition, to

prepare a complete set of values which covers the whole range of conditions under consideration. The 'nuclear data' consists of many kinds of physical terms : the nuclear reaction cross sections, resonance parameters, energy and angle distributions of emitted particles, radioactivity and fission-product yield data, and so on. This paper deals with the statistical evaluation of cross sections obtained from experiments.

It is a common practice to fit experimental data for the evaluation when there is no suitable theoretical model to estimate the cross section for a certain energy range of an incident particle. The fitting approach is also adopted when there is a need to estimate the model parameters necessary to complete a physical model. There are many general-purpose computer programs to fit data and, in the nuclear field, even over 10 least-square fitting programs are available from OECD/NEA Data Bank[1]. The purpose and applicability of the programs, however, are different each other and seldom meet the requirements for this study. The requirements on the program established for the study are as follows. The variance-covariances of raw data affect the resulting evaluated values themselves and also the confidence level of evaluated ones[2]. In addition, the variance-covariances of evaluated quantity are requested in ENDF-6 format and their importance has been emphasized since the early 1970's [3(papers 4.1 and 4.2), 4]. Therefore the program shall treat variance and covariance data which are provided along with the experimental value and shall give variances and covariances of evaluated, i.e. fitted values. As the second requirement, the access to the program shall be easy. Since the program ultimately will be one of the modules in an integrated nuclear data evaluation code package, the well-defined interface and accessibility to the source program are essential. The third one is on the supplementary output items. Since the fitted result by itself is not acceptable sometimes, the evaluator always should review the result and decide to accept or not. Other information, *e.g.* the goodness of fit, is helpful to the decision and the program shall provide such information additional to the fitted values and their variance-covariances.

A computer program COVAFIT which meets above requirements has been developed. COVAFIT fits experimental reaction cross section against the energy of incident particle. It adopts the weighted lin ear least-square regression method with high-order polynomials. Section 2 describes the method including the newly derived orthogonal polynomial. One of the major features of COVAFIT is the capability to treat the variance-covariance matrix provided along with the raw cross section data. Detailed features are found in subsection 2.2. Section 3 shows the evaluat ed neutron total cross sections of oxygen and proton interaction cross section for a C^{11} production reaction with discussions. The final section is devoted to conclusions and suggestions for the further improvements.

2. Fitting Methodology and Features of COVAFIT

2.1. Weighted Least-Square Regression Methodology

A function y(x) is presented by following equation in COVAFIT.

$$y(x) = b_0 P_0(x) + b_1 P_1(x) + \cdots + b_{K-1} P_{K-1}(x) , \quad (1)$$

where the coefficients b's are determined through the regression. This equation is linear to the regression parameter vector **b** regardless of the form of function vector **P**(x). The function **P**(x) will be derived later in this subsection.

To begin with, suppose the model characterizing N experimental data :

$$\mathbf{Y} = \mathbf{X}\,\beta + \varepsilon ,$$

where **Y** is the vector of observed values $(y_1\ y_2 \cdots y_N)^t$, **X** the N by K design (or sensitivity) matrix such as

$$\mathbf{X} = \begin{pmatrix} P_0(x_1) & P_1(x_1) & \cdots & P_{K-1}(x_1) \\ P_0(x_2) & P_1(x_2) & \cdots & P_{K-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ P_0(x_N) & P_1(x_N) & \cdots & P_{K-1}(x_N) \end{pmatrix},$$

and ε is the error vector of which the expected value, $E(\varepsilon)$, equals to null vector and the variance-covariance matrix, $V(\varepsilon)$, equals to $V\sigma^2$. The off-diagonal elements of $V(\varepsilon)$ may not be zero because of correlated observations. The governing equation, *viz.* normal equation, for the above model is

$$\mathbf{X}^t \mathbf{V}^{-1} \mathbf{X} \mathbf{b} = \mathbf{X}^t \mathbf{V}^{-1} \mathbf{Y}$$

where the vector $\mathbf{b}$ is the estimation of $\boldsymbol{\beta}$[5]. Then the solution becomes

$$\mathbf{b} = (\mathbf{X}^t \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{V}^{-1} \mathbf{Y} \qquad (2)$$

and its variance-covariance matrix is

$$\mathbf{V}(\mathbf{b}) = (\mathbf{X}^t \mathbf{V}^{-1} \mathbf{X})^{-1} \sigma^2 . \qquad (3)$$

The regression parameter vector $\mathbf{b}$ is an unbiased estimation of $\boldsymbol{\beta}$ if the design matrix has been chosen adequately. The inverse of $\mathbf{V}$ in Eqs.(2) and (3) is a kind of weighting matrix and the inversion is relatively easy because $\mathbf{V}$ is symmetric and positive definite normally. Cholesky's method [6] is adopted in COVAFIT to get the inverse. If only the observations of a physical quantity under consideration are reported without their variances and covariances, $\mathbf{V}$ becomes unit matrix and $2\sigma^2$ in Eq.(3) is assumed to be equal to the mean square errors between the fitted and given raw data.

Now let us concentrate on how to construct the design matrix $\mathbf{X}$. It is required in Eqs.(2) and (3) to inverse the matrix $\mathbf{X}^t \mathbf{V}^{-1}\mathbf{X}$. Even the inversion is not so difficult because the matrix size is relatively small, only K by K, there are advantages attributed to the diagonalization of the matrix. Additional to the reduced computational effort in the inversion another advantage is as follows. The diagonalization results in zero covariances between regression parameters, b_i's, in Eq.(3). Then the values of parameters are not altered by the change of fitting order K, and only parameters corresponding to the higher terms are re-computed when K is increased. This characteristics is indispensable to find a statistically optimal fitting order without excess computational effort.

The (k, l) element of $\mathbf{X}^t \mathbf{V}^{-1}\mathbf{X}$ is written by

$$\sum_{i=1}^{N} \{P_k(x_i) \sum_{j=1}^{N} w_{ji} P_l(x_j)\},$$

where ω_{ji} is the (j, i) element of $\mathbf{V}^{-1}$. Following recursion relation, which is derived in this study, makes the off-diagonal terms $(k \neq l)$ be zero :

$$P_{k+1}(x) = (x - \gamma_k) P_k(x) - \sum_{l=0}^{k-1} C_l^k P_l(x) ,$$

where

$$\gamma_k = \frac{\sum_{i=1}^{N} x_i P_k(x_i) \sum_{j=1}^{N} w_{ji} P_k(x_j)}{\sum_{i=1}^{N} P_k(x_i) \sum_{j=1}^{N} w_{ji} P_k(x_j)},$$

$$C_l^k = \frac{\sum_{i=1}^{N} x_i P_k(x_i) \sum_{j=1}^{N} w_{ji} P_l(x_j)}{\sum_{i=1}^{N} P_l(x_i) \sum_{j=1}^{N} w_{ji} P_l(x_j)},$$

and $P_0(x) = 1$. The orthogonality achieved by applying the above relation eliminates, at least in principle, the so-called multicollinearity and assures non-singularity of $\mathbf{X}^t \mathbf{V}^{-1}\mathbf{X}$.

Since the ANOVA (ANalysis Of VAriance) is essential to any fitting program it is described briefly here rather than in next subsection. The ANOVA provides a measure to check the goodness of fitted results. The F-test is utilized in COVAFIT to judge whether the regression is significant or not and how much each term in Eq.(1) is significant. The statistical terms, sum of squares due to residual errors (SSE) and due to regression (SSR), are calculated by

$$SSE = (\mathbf{Y} - \hat{\mathbf{Y}})^t \mathbf{V}^{-1} (\mathbf{Y} - \hat{\mathbf{Y}}) ,$$

$$SSR = \sum_{i=1}^{K-1} SS(b_i) ,$$

$$SS(b_i) = b_i^2 (P_i(x_1)\ P_i(x_2) \cdots P_i(x_N))$$
$$\mathbf{V}^{-1} (P_i(x_1)\ P_i(x_2) \cdots P_i(x_N))^t ,$$

where $\mathbf{Y}$ is the raw data vector and $\hat{\mathbf{Y}}$ the fitted vector. The F_0 value for the regression is equal to (SSR/SSE). (N-K-2)/(K-1) and this value is compared to that of $F(K-1, N-K-2; \alpha)$ with the level of significance α. The significance of each parameter b_i is calculated similarly and the contribution of i-th term on the regression is judged. Meanwhile the covariance between two fitted values at x_a and x_b is calculated by

$$\mathrm{Cov}\,(\hat{y}_a, \hat{y}_b) = (P_0(x_a) \cdots P_{K-1}(x_a))\, \mathrm{V(b)}\, (P_0(x_b) \cdots P_{K-1}(x_b))^t$$

and this becomes the variance if $x_a = x_b$.

2.2. Features of COVAFIT

The FORTRAN 77 program COVAFIT runs at an IBM compatible PC. The input consists of direct user input and a raw data file. An EXFOR file[2] which contains experimental nuclear data under evaluation is directly used as the input. EXFOR files contain various kind of data from file to file and are distributed by OECD/NEA Data Bank. The output consists of fitted values and their variance-covariances, the measures to show the significance of the regression, and other information such as followings.

The optimal fitting order recommended in COVAFIT is determined by comparing the F_0 values for fitting order of from 1 to K-1. Owing to the orthogonality no additional computation effort to the regression parameters is needed for different orders.

Identifying outliers is one of the features of COVAFIT. An outlier is a datum which deviates far from the trend of other observed values. It is normally due to an experimental fault, however in cross sections, the resonance region physically includes outlying data in the sense of data fitting. The outliers affect the results to a great extent and it is required to treat them separately if those are valid physically. In COVAFIT the standardized residual at each energy point is calculated and, if the value is greater than the critical value by Lund[7], it is noticed as an outlier.

It is restricted to apply COVAFIT to the cases of small or infrequent variations in cross sections. So multiple piecewise COVAFIT fittings and/or post-COVAFIT processing are recommended to cover a very wide energy range and steep variations in cross sections. COVAFIT yields knots for the spline interpolation which is one of the post-processing tools. Values of the regression function, its derivative, and the second derivative become zero at the recommended knots.

3. Applications and Discussions

3.1. Evaluation of Neutron Cross Section of Oxygen

Total cross sections of both O^{16} and natural oxygen for MeV range neutrons have been evaluated. EXFOR entry 10047 is used for O^{16} cross section evaluation and EXFOR entry 20742 for natural oxygen. EXFOR 10047 (Compiled Reference : D.G. Foster Jr. and D.W. Glasgow, 1971) provides both variances and penta-diagonal covariances in addition to the cross sections for over 240 energy points covering from 2.5 to 15 MeV. Meanwhile EXFOR entry 20742 (Compiled References : S. Cierjacks *et al.*, 1980, etc.) contains total cross sections of over 21,000 points for 3.1 to 32 MeV incident neutrons, and it was used in JENDL-3 for higher than 3 MeV neutrons[8].

Fig. 1 shows the evaluated O^{16} total cross section. The whole energy range is divided into 7 intervals, as indicated by vertical dotted lines in the figure, and fitting order varies from 5 to 17 for each interval. Although there is, in general, no direct measure to justify the final acceptance of the fitted result except the evaluator's decision, this figure looks well-evalu-

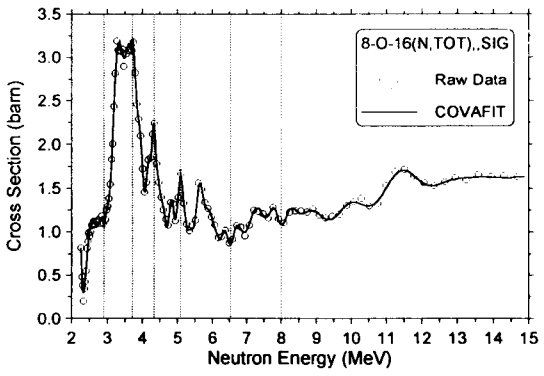

Fig. 1. Evaluated Neutron Total Cross Section of O^{16}

ated and the program COVAFIT is satisfactory in its function. An issue, however, arises : how to divide the whole energy range into small intervals. It is the conclusion that the energies having peak or dip cross section are taken as the interval boundaries since the regression tends to yield underestimated cross sections for peak portions or overestimated ones for dips within an interval. Fig. 2 shows an example of the behavior of standard deviation, which equals to square-root of variance, of estimated cross sections for 2.9~3.7 MeV region. The raw data in this figure means the reported experimental standard deviations. The larger values near interval boundaries are statistically natural; in general, the loci of confidence limits of a regression show the form similar to a concave lens. However it is not true physically in this case because the boundaries are ones determined arbitrarily by the evaluator. Therefore an overlap between intervals is necessary to remove such unphysical large variances. For example, let us consider six energy points ovelapping case which means six points are included in both adjacent intervals. Two estimated variances are obtained for each overlapped point, then the variances at three energy points near to the boundary are discarded. Overlapping four to ten experimental energy points is adequate to obtain a set of meaningful variances through the whole energy range. Fig. 2 also shows the difference caused by whether the raw standard deviations are utilized or not. The solid line is for the case accounting for given variance-covariances data and the dotted line for the case without those data. The estimated cross sections of the two cases are very close each other, but the resulting variances are different. On the other hand, the estimated covariances show a correlation between any two estimated cross sections in an interval and it is observed that there are negative covariances with small magnitudes.

COVAFIT has been applied also to the evaluation of total cross section of natural oxygen with EXFOR entry 20742. The results confirm again the applicability of the program for smooth varying cross sections, higher than 7 MeV in this case. The program works well as shown in Fig. 3 even for a sharp resonance. However it reveals a weakness for consecutive sharp resonances in one energy interval. Although the weakness can be overcome by taking narrower intervals, it needs more evaluation effort. Increasing the order of polynomial is the other solution but it decreases the statistical significance of the regression. So a pre-processing of the raw data has been considered, but not accommodated yet, in the following way. The pre-processing flattens the data by subtracting the resonance portion, which is calculated by applying certain resonance model, from the raw data. Then it is added later to the evaluated result obtain-

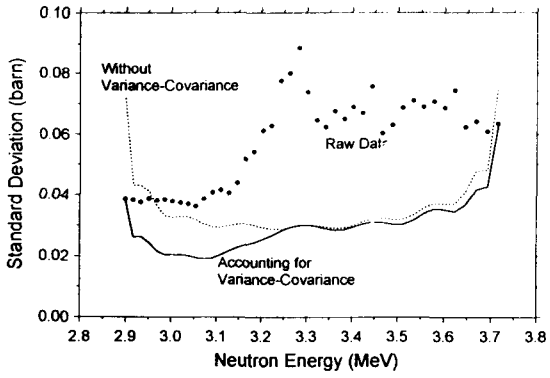

Fig. 2. An Example of Standard Deviations in Evaluated Cross Sections

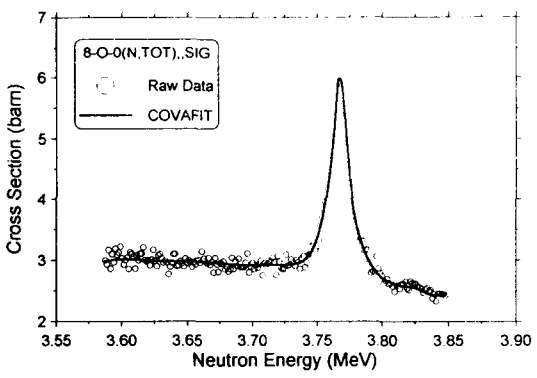

Fig. 3. An Example of Evaluated Cross Section around a Resolved Resonance

ed with flattened data. It seems that a rather crude resonance model is sufficient, but the treatment of variance-covariance remains as an issue.

3.2. Evaluation of Proton Cross Section for C^{11} Production Reaction

The positron (e^+) emission tomography (PET), one of the advanced medical diagnosis equipments, uses radioactive isotopes such as C^{11}, N^{13} and O^{18} to get *in vivo* information[9]. The proton interaction cross sections for three C^{11} production reactions, $B^{11}(p, n)C^{11}$, $C^{12}(p, p+n)C^{11}$ and $N^{14}(p, \alpha)C^{11}$, have been evaluated. Since there are, in essence, no remarkable differences in the results for all three cases, only the evaluation for the first reaction is reported in this paper.

Four EXFOR entries, A0330(B. Anders, 1981), B0106(G. Albouy, 1962), P0001(J.H. Gibbons, 1959), and P0045(M. Furukawa, 1960), have been merged into one file before applying COVAFIT, and the energy scale has been transformed by taking natural logarithm because of the broad proton energy range. Fig. 4 shows the resulting cross section curve. It is worthy to note that the magnitudes of standard deviations in raw data have a broad range, 2.5 to 25% of cross section from file to file and from energy to energy (Error bars are not presented in the figure for legibility). Then the key is whether COVAFIT reflects the importance of each experimental datum, and it seems that the program has done. The evaluated cross sections for the above-mentioned other two C^{11} production reactions also show the applicability of the program.

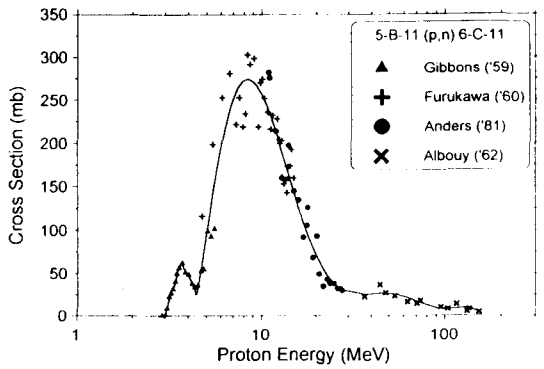

Fig. 4. Evaluated Cross Section of $B^{11}(p, n)C^{11}$ Reaction

4. Conclusions and Recommendations

The program COVAFIT yields meaningful evaluated cross sections and their variance-covariances in all five cases attempted. Including especially the newly proposed method to construct the design matrix, mathematical formulations are adequate enough and the program meets the requirements mentioned in Sec. 1. Since there is no measure to confirm the validity of the produced covariances[3(paper 1.1)], the discussions in the paper are focused on the evaluated cross sections themselves. Even so it shall not be overlooked that the most important feature of COVAFIT is the capability to treat the covariances of raw data.

The program, however, requires much time and effort to evaluate sharply and highly fluctuating data. It is expected that a special treatment for resolved resonances can reduce the burden. Meanwhile adopting the so-called segmented regression method can also do, but the problem may become non-linear to take the best of merits of the method.

An enhanced version might include advanced features. Features recommendable are as follows: a graphic output interface for the easy eye checking, enhanced input/output interface for multiple evaluations in a run, and spline interpolation routine as one of the choices.

Acknowledgement

The authors are indebted to Jong Chan Kim, Department of Physics, Seoul National University. He and his graduates collected and reviewed about fifty EXFOR files on C^{11} production reactions for the study. We are grateful also to Jung Do Kim, KAERI, for his valuable comments on the work.

References

1. Nuclear Program Abstracts, OECD/NEA Data Bank, Jan (1993)
2. 장종화 등, "핵자료 평가 체제 구축 (II)," KAERI/RR-1535/94, KAERI (1995)
3. T. Kawano, Ed., "Covariance Matrix Calculated from Nuclear Models," Proc. of the Specialists' Mtg. on Covariance Data, JAERI-M 94-068, March (1994)
4. D.L. Smith, "Probability, Statistics, and Data Uncertainties in Nuclear Science and Technology," OECD/NEA Nuclear Data Committee Series, Vol. 4, Chapter 13, American Nuclear Society (1991)
5. N.R. Draper and H. Smith, "Applied Regression Analysis," 2nd Ed., Chapter 2, John Wiley & Sons, Inc., New York (1981)
6. 김창효, "수치해법과 전산 프로그래밍," 제6장 제2절, 교학사, 서울 (1984)
7. 박성현, "회귀분석," 개정판, 제16장 제2절, 민영사, 서울 (1991)
8. K. Shibata, et al., "Evaluation of Neutron Nuclear Data for ^{16}O," JAERI-M 90-012, JAERI, Feb (1990)
9. 조규성, "PET의 개발 현황," 제1회 의료용 가속기 학술회의-Cyclon과 그 응용, 원자력병원, 서울 (1994)