объединенный
институт
ядерных
исследований
дубна

E5-88-463 e

A.Blagoev*, T.Mishonov, S.Kovatchev*,
N.Pilosof*

BOOTSTRAP AND JACKKNIFE SOLVING OF LINEAR EQUATION SYSTEM FOR EXPERIMENTAL DATA PROCESSING

Submitted to "Journal of Physics A"

*Faculty of Physics, Sofia University, Bulgaria

1988

INTRODUCTION

A great number of data processing problems leads to the solving of the overdetermined linear equation systems. Well developed methods and standard computer routines exist for the solution of similar systems. Most of them are based on the least squares method proposed by Legendre and Gauss. As a rule, these methods give satisfactory solutions of the equation systems, but the information about the dispersion and the confidence intervals is insufficient. More information for the quality of the obtained solution gives the SVD (singular value decomposition) by which the number of the well-established variables could be estimated [1]. In most cases it is possible to evaluate the accuracy of the obtained solutions in general, but not the accuracy of each one variable.

Since 1950 a series of nonparametric methods, named Jackknife and bootstrap [2,3] has appeared in the statistics , which affords a possibility for a new approach towards data processing. Roughly speaking, the basic idea of this approach is the multiplication of a limited number experimental data in a substantially bigger assembly, using a considerable computer recourses (mainly calculation time). The essential advantage of these methods is the freedom from normality assumption. The development of the above methods is stimulated by the spreading of the fast computers [4,5]. In many cases it is cheaper nowadays to increase the quantity of computation instead of increasing the volume of the stored empirical information.

In the present work the ideas of these methods are applied in solving and estimating of the accuracy of the obtained solutions of overdetermined system, resulting from the processing of the data yield by the glow discharge experiments. We have to use this approach because the obtained values of various rate constants differ approximately one order of magnitude; then arises a question about the accuracy of each of the values determined in this way.

EXPERIMENT

The number densities of the $2p^5 3s$ $^3P_{2,1,0}$ excited neon atoms are measured by optical absorption during the afterglow of the neon positive column. Simultaneously, the number density and the temperature of the electrons are measured by electrical probes. More extensive description of the experimental set up and the obtained results are given in [6]. The balance equation for the metastable Ne^3P_2 state at the specific experimental conditions of this work is:

$$\frac{d(\ln N_2)/dt-\gamma_2}{N_e}=K_{21}\left[\frac{N_1}{N_2}\,\frac{5}{3}\,\exp\frac{\Delta E_{21}}{kT_e}\;-1\right]+K_{20}\left[\frac{N_0}{N_2}\;5\;\exp\frac{\Delta E_{20}}{kT_e}\;-1\right], \quad (1)$$

where $N_{2,1,0}$ are the number densities of neon atoms $Ne3^3P_{2,1,0}$, ΔE_{ij} are the energy gaps between the $Ne3^3P_i$ and $Ne3^3P_j$ levels, γ_2 is a "pressure effects" term; N_e and T_e are the electron density and temperature, K_{ij} are the rate constants of the reactions $Ne3^3P_i \xrightarrow{e} Ne3^3P_j$. It is shown in [6] that the temperature dependance of the rate constants is $K_{ij}=k_{ij}\exp(-\Delta E_{ij}/kT_e)$ in the region 400-4000 K. Substitution for these dependencies in eq.(1) gives a linear equation for the constants k_{21} and k_{20}. Measuring the corresponding values in different moments of the plasma decay and various discharge conditions (current and gas pressure), we get a heavily overdetermined system of equations in the form of equation (1). In our case we have 126 equations for two variables.

METHOD

The overdetermined system of M linear equations for N unknown quantities is solved using the least squares method (LSM) by reducing it to a determined system of the so-called normal equations. In a matrix form if the initial system is $Ax = b$, the normal system is $(A^TA)x = (A^Tb)$. Here is the matrix ' $M \times N$) of the coefficients, x is the vector $(x_1,\dots,x_N)$ of the solution and b is the free terms vector.

The Jackknife method for statistical processing of the temporal series is proposed by M.Quenouille [7] and J.Tukey [8]. The modification of the above method, which we use to solve the undetermined system of M linear equations with N variables is as follows : from all M equations, J equations are selected ($N \le J \le M$) by arbitrary choice. This new system of J equations is solved by the LSM. In this way $\begin{pmatrix} J \\ M \end{pmatrix}$ different subsystems of the source system are derived which have in general different solutions. Repeating many times this procedure of arbitrary selection and solving J equations subsystems we have a large population of approximate solutions of the initial system. Analysing the distributions of these solutions a conclusion about the obtained values accuracy could be made. Using the method of percentiles to characterize the confidence intervals as Efron has done in [9] we choose 25% and 75% as lower and upper limits. The difference between these percentiles is the full width at the half maximum (FWHM) of the distribution.

In the Bootstrap method by means of N random samplings with replacement of equation, a system similarly to the initial one is formed and then solved by the LSM. A manyfold repeating of this procedure (> 1000 times) gives the bootstrap distribution of the solutions. In this particular case the Bootstrap procedure is equivalent to a multiplying of the initial system equation by a non-negative integer weight coefficient, the total amount of which is equal to the number of equations M. Besides this classical version of the Bootstrap algorithm two other its modifications were applied. One natural weak generalization of the Bootstrap is to use uniform distributed "real" numbers, produced by random generator instead of the integer weight coefficients. More deep modification of the Bootstrap turns to be a multiplication of the initial equations by a random number and their summation. Repeating this procedures much as N times, where N is the number of variables, a determined system is created. This system could be solved by standard methods. Thus, a large aggregate of the approximate solutions to the initial system could be created. In a matrix form this modification of Bootstrap could be written similarly to LSM as $(RA)x = (Rb)$, where R is the matrix of random numbers with the dimension of A^T.

The meaning sense of the application of such procedures clears up by the following consideration : When experimental data are processed by the LSM it is presumed that all data are equally reliable or that the data quality could be taken into

account by the introducing weight coefficients. However, in the typical experimental situation it is quite difficult to estimate the relative quality of different measurements, especially when the link between the experimental numbers and the values which should be determined by data processing is a complicated one as it is in the solving of linear equation system. The application of the Bootstrap in this case could be regarded as a test of variety of assumptions for the relative quality of the separate equations by introducing arbitrary coefficients.

RESULTS

The distributions of the solutions obtained by 5000 trial with the Bootstrap method for the first and the second variable respectively are shown on Fig.1a and 1b. The median and the corresponding percentiles (25% and 75%) are also shown. The values of the variables are normalized on the real values of the rate constants published in [6] which has been obtained using LSM. On Fig.2a and 2b are shown several Jackknife-distributions at different J values. Figures 3a and 3b show the plots of the median and the corresponding percentiles versus the number of the selected equations J. The Table shows the values of medians, percentiles and FWHM for the two variables obtained by different methods. For the LSM the 50% confidence interval - $2z(0.25)\sqrt{\sigma^2}$ is used as FWHM. Here $z(0.25)=0.67$ is 25% point of the standard normal distribution and σ^2 is the estimation for the dispersion in linear regression model.

It could be seen that the different methods give substantial difference in the estimations of the errors. The estimations of the results accuracy yield by the methods designed in the Table as II and VII are too pessimistic. Besides that the medians of the corresponding distributions are quite

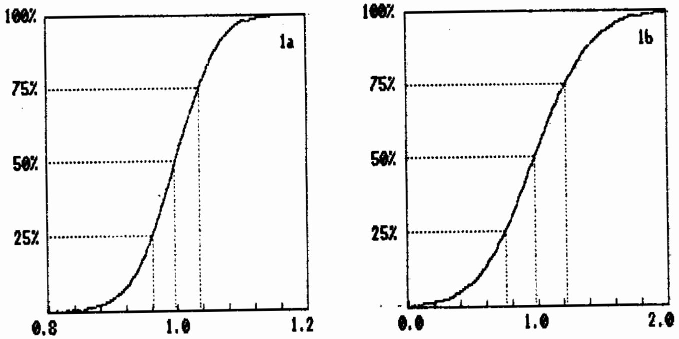

Fig. 1 Bootstrap distributions with the percentiles (5000 trials)

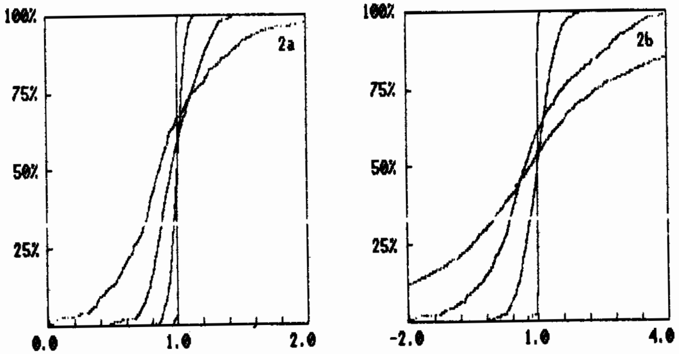

Fig. 2 Jackknife distributions at J=2,10,63 and 125 (M=126; 1000 trials)

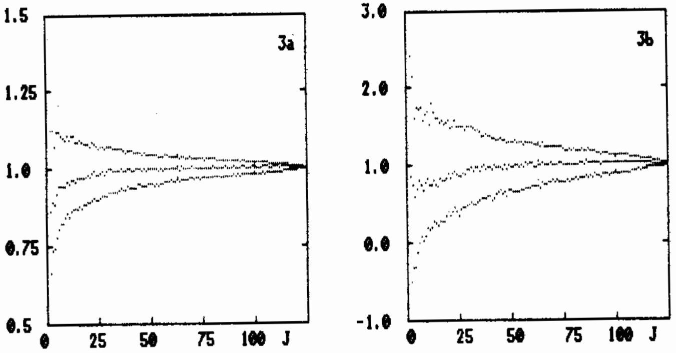

Fig. 3 Jackknife distributions percentiles versus number of selected equations J

Table 1

Method	$\bar{x}_1$	$x_1^l - \bar{x}_1$	$x_1^u - \bar{x}_1$	$FWHM_1$	$\bar{x}_2$	$x_2^l - \bar{x}_2$	$x_2^u - \bar{x}_2$	$FWHM_2$	$\frac{FWHM_2}{FWHM_1}$
I	1.00	-	-	0.040	1.00	-	-	0.37	9.3
II	0.86	-0.21	+0.26	0.47	0.84	-1.34	+1.54	2.88	6.1
III	1.00	-0.038	+0.041	0.079	0.99	-0.27	+0.26	0.53	6.7
IV	1.00	-.0002	+.0008	.0009	1.00	-0.006	+0.009	0.014	15.5
V	1.00	-0.037	+0.037	0.074	0.98	-0.22	+0.24	0.46	6.2
VI	1.00	-0.036	+0.035	0.071	1.00	-0.21	+0.23	0.44	6.2
VII	0.96	-0.22	+0.23	0.45	1.05	-2.51	+2.66	5.17	11.5

Methods : I - Least squares; II - Jackknife (J=N; N - number of variables); III - Jackknife (J=M/2; M - number of equations); IV - Jackknife (J=M-1; "standard"); V - Bootstrap (standard - selection with returning); VI - Bootstrap (small modification - multiplying by real random coefficients); VII Bootstrap (modified - multiplying by random matrix of coefficients)

$\bar{x}$ - medians of the solution distributions (50% percentile);

x^l - 25% percentile(low); x^u- 75% percentile(up);

FWHM = (x^u-x^l) - Full Width at Half Maximum;

unlike to the solution given by the LSM (method I). On the contrary IV method gives too optimistic estimation. The estimations given by III,V and VI methods correlates quite well, but differ substantially from the one yield by the LSM. The ratio of the FWHM of both the variables shown in the last column also differ substantially. In our opinion one consideration supporting the estimations given by methods III,V and VI is that the ratio of the two variables errors is near to the ratio of the variables themselves written in the same units.

In this way the application of the Bootstrap and Jackknife methods permits one to abjust the degree of confidence by which every variable is determined.

CONCLUSIONS

Our practice in using of the Bootstrap and Jackknife methods shows that they could be a quite useful instrument in the processing of the experimental results. The application of these methods could give in some cases an indication for incorrectness of the used model (for example when strongly asymmetrical distributions appeared).

ACKNOWLEDGMENTS

The authors are grateful to Dr. Tc. Sarijski who drew our attention towards these methods and to Dr. P. Fiziev for useful discussions.

REFERENCES

1. Forsythe J. E, Malcolm M A, Moler C B, 'Computer Methods for mathematical computations, Prentice-Hall,N.J.,1977
2. Miller R.G, 1974,Biometrika,61,1-17
3. Efron B, 1979, Ann.of Statistics,7,1-26
4. Efron B, 1979, SIAM Review,21,460-480
5. Diaconis P, Efron B,1983,Scientific American,248,N5
6. Pilosof N, Blagoev A, 1988, J.Phys.B:At.Mol.Opt.Phys.,21, 639-646
7. Quenouille M.H, 1956,Biometrica,43,353-360
8. Tukey J.W, 1958, Ann.Math.Statist.,29,614
9. Efron B, The Jackknife, the Bootstrap and Other Resampling plans , SIAM , Philadelphia , 1982

Received by Publishing Department
on June 29, 1988.

Благоев А. и др. E5-88-463
Бутстреп метод и метод складного ножа
для решения линейных систем уравнений
для обработки экспериментальных данных

Обсуждается применение нетрадиционных методов многомерного статистического анализа для решения переопределенных систем линейных уравнений; методы иллюстрируются на примере исследования элементарных процессов в послераспадающейся плазме.

Работа выполнена в Лаборатории теоретической физики ОИЯИ.

Препринт Объединенного института ядерных исследований. Дубна 1988

Blagoev A. et al. E5-88-463
Bootstrap and Jackknife Solving
of Linear Equation System for Experimental
Data Processing

The application of statistical methods known as bootstrap and Jackknife is considered for solving the overdetermined linear equation system, which is illustrated by an example concerning elementary processes in the afterglow plasma.

The investigation has been performed at the Laboratory of Theoretical Physics, JINR.

Preprint of the Joint Institute for Nuclear Research. Dubna 1988