Available online at www.sciencedirect.com
SCIENCE DIRECT®
Nuclear Instruments and Methods in Physics Research A 502 (2003) 827–828
NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A
www.elsevier.com/locate/nima

Summary talk—V Section

E.L. Kosarev

P.L. Kapitza Institute for Physical Problems, Russian Academy of Sciences, ul. Kosygina 2, Moscow 117334, Russia

Abstract

This talk presents a summary of the reports included in Section V of this Workshop.

PACS: 07.05.Kf; 06.20.Dk; 02.50.−r

Keywords: Deconvolution; Classification; Histogramming

There are several groups of reports on "Advanced Statistical Methods for Data Analysis":

- Methods, which have significantly used various deconvolution methods: reports of V. Anikeev, G. Chabratova, E. Kosarev, and M. Morhac.
- Different methods for physical events classification and separation: reports of D. Anipko, Y. Kulchitsky, J. de Seixas, A. Vardanyan, P. Temnikov, and L. Karimova.
- Different methods for the estimation of accuracy of various physics parameters obtained from the measurements: reports of S. Redin, S. Alekhin, S. Manayenkov, D. Anipko, F. Tkachev, and S. Bityukov.
- One of the very interesting reports by V. Ivanov using methods mainly from the second group is demonstrated by the statistical properties of network traffic. It was shown in this report, probably at a first time, that the package size distribution has the log-normal distribution.

Because I cannot expound more information than the authors of the reports I will only discuss some points from each of the above report groups.

- Only one nonlinear method for deconvolution—maximum entropy is mentioned in the report of V. Anikeev, but it has not mentioned the MLM. Though the results of both methods are usually very close to one another, the statistical meaning and computer programs can be very much distinguished from one another. It would be useful for G. Chabratova and for future readers of her report to compare the obtained results with the previously published. There is no substantiation in the vast report of M. Morchac (and also in the cited paper NIMA 401 (1997) 113–132) on the method of background elimination and also the general structure of working algorithm: background elimination, than unfolding and fitting.
- It is demonstrated in the two reports of Y. Kulchitsky and of Jose de Seixas the various approaches of data treatment for the same set

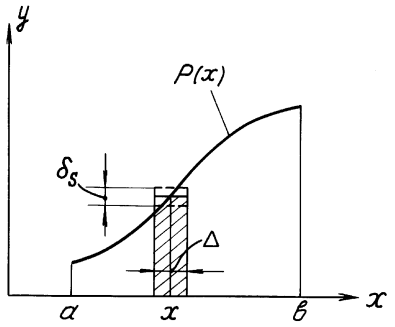
E-mail address: kosarev@kapitza.ras.ru (E.L. Kosarev).

0168-9002/03/$ - see front matter
doi:10.1016/S0168-9002(03)00611-9

of HEP detectors. It would be interesting to compare both results together. Also two reports with application to cosmic physics (of A. Vardanyan and of P. Temnikov) have demonstrated the very impression results obtained by using the Neural Net approach. In the interesting report of L. Karimova is illuminated the different methods of random data classification and analysis and also the very different examples of the real data.

- The report of S. Redin deals with the old story problem in mathematical statistics: what method is better—min χ^2 or maximum likelihood, and how to estimate the unknown probability density function? A vast literature both in textbooks and monographs and also in journals is devoted to this subject. A best way to know about this problem is to find any of the new paper on this subject and to follow the consecutive references from this paper to the previous ones. The report of S. Bityukov has demonstrated one possibility of the definition of weak signal registration significance. As concerned the report of S. Manayenkov it would be useful for him to note the two References: (a) F. James and M. Roos. Nucl. Phys. B 172 (1980) 475 and (b) L.N. Bolshev, N.V. Smirnov The Tables of Mathematical Statistics, Moscow, 1968 (in Russian)—estimation of the ratio of two Possonian variables (Table 5.5).

Addendum: A remark on optimal histogram binning



Let Δ be the width of the histogram bin and $p(x)$ the sought for unknown probability density. There are two source of errors by histogram approximating of $p(x)$: the statistical one

$$\delta_s = \frac{1}{\sqrt{Np(x)\cdot\Delta}}$$

and the systematical one

$$\delta_a = p'\cdot\frac{\Delta}{2}/p.$$

Total accuracy is

$$\delta^2 = \delta_s^2 + \delta_a^2 = \frac{1}{Np\Delta} + \frac{\Delta^2}{4}\left(\frac{p'}{p}\right)^2.$$

Looking min δ^2 as a function of Δ,

$$\Delta_{\text{opt}} = \left(\frac{2p}{Np'^2}\right)^{1/3},$$

$$\sqrt{(\delta^2)_{\min}} = \frac{\sqrt{3}}{\sqrt[3]{4}}\frac{1}{N^{1/3}}\left(\frac{p'}{p^2}\right)^{1/3}.$$

The optimal number of histogram bins are

$$N_{\text{opt}} \sim N^{1/3}.$$

This is the result of N.V. Smirnov: Soviet Math. Doklady AN SSSR, 74, (2) (1950) 189 (in Russian).