объединенный
институт
ядерных
исследований
дубна

E11-87-367

B.Słowiński

SIMPLE STATISTICAL MULTIVARIATE APPROACH TO WEAK SIGNALS EXTRACTION

Submitted to the Second International Tampere Conference in Statistics, June 1-4, 1987, Finland.

1987

1. INTRODUCTION

Numerous methods of applied statistics are widely used in many fields of knowledge and practice. In high energy physics where experimental devices and their operation are especially expensive, convenient selection of the most effective approaches to analysing data is very important. On the other hand, a discussion of concrete examples being drawn from different domains and illustrating great necessity of statistical treating of empirical data can stimulate in the first place the development of such methods which have direct practical significance.

The paper contains two examples of simple application of some elements of the multivariate analysis. In general they concerns a problem of a detection of weak signals accompanied by a substantial background.

Let us consider a situation, typical enough in experimental high energy physics, when heavy relativistic fragments (protons, deutrons, tritons, alphas) emerging from inclusive nuclear reactions are detected by an electronic device (see, for example, /1-3/). If alphas are used as impinging particles and light nuclei as a target, then protons and deutrons are predominantly emitted particles. So, heavier secondary fragments which are of great interest too, produce very small signals and their indentification is a problem equally important as difficult.

Relativistic fragments can be in principle identified by their electrical charge Z (latter: charge) and rest mass M(latter: mass). Information about a charge can be drawn from the ionization effect registered using scintillation counters (SC) as a random signal of the amplitude $A_j \sim Z^2$ /4/. Hence for each Z (being equal to 1 or 2) and each i-th SC we have the (Gaussian-like) distribution $f_{Z,i}(A_j)$. Furthermore the particle mass M is easy to estimate by means of the measuring of time-of-flight of the particle penetrating through the k-th couple of counters. Then we have again, as above, the (Gaussian-like) distribution $f_{M,k}(M_j)$ for particles of a given kind. Remark must be made yet that the mass M is to estimate after the particle charge Z is established only, because in this case at each Z value there are different M_j - distributions. Our goal is to determine from sample the composition function $\sigma(Z,M)$ of secondary particles when in each j-th event (i.e.

for each particle) are detected n independent random numbers $A_j^{(i)}$ ($i = 1,\ldots,n$) and some signals allowing us to get l independent as well random numbers $M_j^{(k)}$ ($k = 1,\ldots,l$) if the particle charge Z is found out earlier. Numerical analysis has been performed using experimental data obtained by means of the MASPIK spectrometer of JINR[3], where n = 5 and l = 2. As follows from the above discussion the problem under consideration may be solved by the two-step method: 1) charge determination, and 2) mass determination.

2. CHARGE DETERMINATION

As has been pointed out previously in each j-th event 5 amplitudes $A_j^{(i)}$ are registered for a particle having the charge Z and the mass M. Because predominantly light component is created in the reaction of alpha particles with light nuclei at 4.5 GeV/N then $A_j^{(i)}$ - distributions, experimentally obtained, correspond practically to one-charge particles, i.e. those having Z = 1. Similar distributions for Z = 2 one can get by different ways, but simplest one and correct enough is to produce them from those at Z = 1 taking into account that $A_j^{(i)} \sim Z^2$. So, in principle it is possible to separate secondary particles by their charge at the acceptable significance level (SL). For this purpose, as usually, it is necessary to choose for each SC a desired value of SL associated with the one-tail test with critical region on the right for the $A_j^{(i)}(Z=1)$ - distributions and to estimate appropriate probabilities of a Type II error (one-tail test with critical region on the left for the $A_j^{(i)}(Z=2)$-distributions). Figure 1 shows the $A_j^{(i)}$ - distributions for all 5 SC obtained at two different conditions of SC operation (solid and dashed histograms). These conditions can be described by means of the variance coefficients (VC) $x_i = \frac{\Delta A^{(i)}}{\bar{A}^{(i)}}$, where $\Delta A^{(i)}$ is the standard deviation and $\bar{A}^{(i)}$ is the average value of the relevent $A_j^{(i)}$ (Z=1) -distribution when random amplitudes $A_j^{(i)}$ satisfying the condition $A_j^{(i)} < 3 \cdot A_o^{(i)}$ are taken into account only ($A_j^{(i)}$ - distribution attains its maximum at the value $A_o^{(i)}$). In Figure 1 $A_j^{(i)}$-distributions on the left correspond to the hypothesis Z=1 and those on the right have been obtained using the condition $A_j^{(i)} \sim Z^2$. Arrows show (for both values of x) three va-

lues of $Q^{(1)}$-quantiles: $Q^{(1)} = 2\bar{A}$, $Q^{(2)} = 2.5\bar{A}$ and $Q^{(3)} = 3\bar{A}$, which are of practical interest and associated with admissible values of probabilities $P_Z^{(i)}$ of both types error ($P_{Z=1}^{(i)} = 10^{-2} \div 10^{-3}$ for a Type I error and $P_{Z=2}^{(i)} = 10^{-1} \div 10^{-4}$ for a Type II error, relevant to each i-th SC and two hypotheses Z=1 or Z=2 correspondingly). Here $\bar{A} = (\bar{A}^{(1)}, \bar{A}^{(2)}, \ldots, \bar{A}^{(5)})$ and $Q^{(1)}$ means therefore 5-tuple criterion as well, i.e. $Q^{(1)} = (Q_1^{(1)}, Q_2^{(1)}, \ldots, Q_5^{(1)})$. So, since all 5 SC are strictly mutually independent we have for the set of these SC $\bar{P}_Z^{(5)} \equiv \prod_{i=1}^{5} P_Z^{(i)}$. Numerical values of $\bar{P}_Z^{(5)}$ for both types error are given in the Table (we stress that a Type I error is associated with the hypothesis Z=1 whilst a Type II error is connected with the alternative hypothesis Z=2).

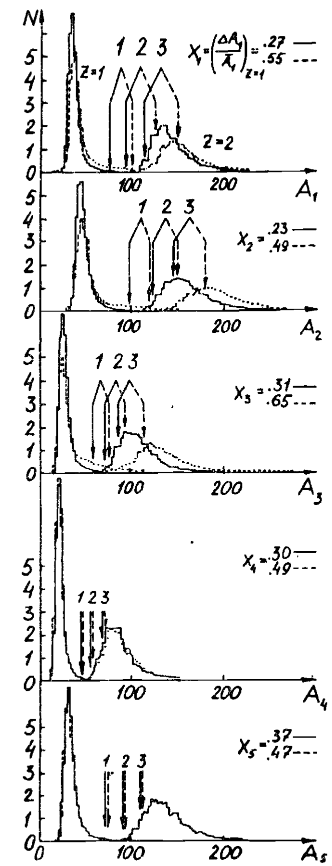

Figure 1

We can see that the 5-tuple criterion only just discussed is very effective one: it makes possible to achieve particle separation by their charge Z if the ratio $r = \sigma(Z=2,M)/\sigma(Z=1,M)$ is such small as about 10^{-14}. Nevertheless, this criterion is sensitive enough with regard to operation conditions, i.e. it markedly depends on x. Therefore it is of interest to consider another combination of five amplitudes $A_j^{(i)}$ as random variables and in the first place the simplest one: $\bar{A}_j = \frac{1}{5}\sum_{i=1}^{5} A_j^{(i)}$ and relevant averaged amplitude criterion. Numerical results for $\bar{A}_j$-distribution and two hypotheses (Z=1 and Z=2) as well as two x coefficient values (x_1= 0.30 and x_2=0.53) are compared in the Table with similar data concerning 5- tuple criterion. One can conclude that although the criterion based on averaged amplitude is more stable with regard to x changing, it is by a factor of even about 10^{10} of magnitude less effective than the 5-tuple one.

Table

Numerical values of the probabilities $\bar{P}_Z^{(5)}$ for all SC and $\bar{P}_Z$ (determined for averaged $\bar{A}_j$ - distribution) associated with $Q^{(1)}$ - quantiles and two alternative hypotheses: Z = 1 and Z = 2. Results are quoted for two samples of former empirical data relevent to different SC operation conditions which are characterized by the variance coefficients x_1 shown in Figure 1. Here $\mathbf{x}^{(k)} = (x_1^{(k)},\ldots,x_5^{(k)})$, $\mathbf{A}_j = (A_j^{(1)}, A_j^{(2)},\ldots, A_j^{(5)})$ and $\mathbf{Q}^{(1)} = (Q_1^{(1)}, Q_2^{(1)},\ldots, Q_5^{(1)})$, each $Q_i^{(1)}$ being equal to: $Q_1^{(1)}=2\bar{A}_1$, $Q_2^{(1)}=2\bar{A}_2$, etc., $\bar{x}_1$=0.30±0.05, $\bar{x}_2$=0.53± 0.06.

P_Z	$\mathbf{x}^{(k)}$	$Q^{(1)} = 2\bar{A}$	$Q^{(2)} = 2.5\bar{A}$	$Q^{(3)} = 3\bar{A}$
$\bar{P}^{(5)}_{Z=1}(\mathbf{A}_j > \mathbf{Q}^{(1)})$	$\mathbf{x}^{(1)}$	$1.1\cdot 10^{-10}$	$0.9\cdot 10^{-12}$	$2.0\cdot 10^{-14}$
	$\mathbf{x}^{(2)}$	$1.1\cdot 10^{-7}$	$6.0\cdot 10^{-9}$	$4.0\cdot 10^{-10}$
$\bar{P}^{(5)}_{Z=2}(\mathbf{A}_j < \mathbf{Q}^{(1)})$	$\mathbf{x}^{(1)}$	$1.0\cdot 10^{-16}$	$1.2\cdot 10^{-11}$	$1.2\cdot 10^{-7}$
	$\mathbf{x}^{(2)}$	$1.0\cdot 10^{-10}$	$3.1\cdot 10^{-7}$	$2.0\cdot 10^{-3}$
$\bar{P}_{Z=1}(\bar{A}_j > \bar{Q}^{(1)})$	$\bar{x}_1$	$(1.1\pm 0.3)\cdot 10^{-2}$	$(4.4\pm 2.3)\cdot 10^{-3}$	$(2.3\pm 1.9)\cdot 10^{-3}$
	$\bar{x}_2$	$(4.5\pm 1.9)\cdot 10^{-2}$	$(2.5\pm 1.0)\cdot 10^{-2}$	$(1.4\pm 0.6)\cdot 10^{-2}$
$\bar{P}_{Z=2}(\bar{A}_j < \bar{Q}^{(1)})$	$\bar{x}_1$	$(2.7\pm 3.6)\cdot 10^{-3}$	$(0.6\pm 0.2)\cdot 10^{-2}$	$(0.9\pm 0.4)\cdot 10^{-1}$
	$\bar{x}_2$	$(1.6\pm 1.8)\cdot 10^{-2}$	$(7.9\pm 7.8)\cdot 10^{-2}$	$(3.1\pm 1.8)\cdot 10^{-1}$

3. MASS DETERMINATION

If the charge Z of a registered particle has been established yet as discussed previously, its mass M can be already estimated correctly. So, we have two values of M (m_1 and m_2) for each particle which are measured independently and defined by charge as well. Then the problem arises again to build a criterion, sufficiently effective and simple at the same time to be used simultaneously when an experiment is in action, which enables us to single out reliably enough the particles being produced with very small probability. It is evident that a criterion based on univariate statistic is too flimsy. To make sure of this let us look at Figure 2 (upper part) where an empiric mass distribution for a sample of size N=4·10^4 of former experimental data is drawn. On the x axis values of m_1 are marked since they are measured with better accuracy than similar m_2 values. We can notice that only protons (p) and deuterons (d) placed within central parts of relevant distributions are to be simply separated in this way whereas other particles (t, alphas and nuclei of ^{3}He) are sinked into complex background originating mainly from long tails of p and d mass distributions. Therefore it is useful to consider a 2-dimensional distribution (or scatter plot) of events consisting of points ($m_1^{(j)}$, $m_2^{(j)}$) whose coordinates are measured values of m_1 and m_2. As an illustration in Figure 3 it is shown the plot of such kind for particles having Z=1. One can see that as expected the majo-

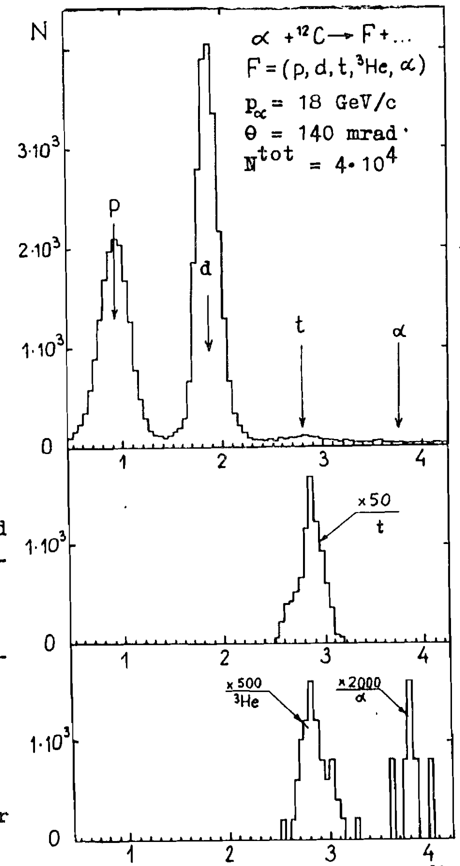

Figure 2 m_1 (GeV/c^2)

rity of points explicitly concentrate within elliptic surface, like the 2-dimensional Gaussian distribution of uncorrelated random variables, i.e.

$$\left(\frac{m_{1i}^{(j)}-M_i}{\sigma_{1i}}\right)^2 + \left(\frac{m_{2i}^{(j)}-M_i}{\sigma_{2i}}\right)^2 \le p^2, \quad (1)$$

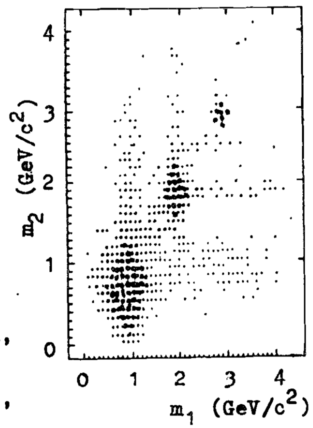

Figure 3

where M_i means exact value of the mass of particles of i-th sort, σ_{1i} (σ_{2i}) is the standard deviation of the central part of associated m_{1i}(m_{2i})-distribution, i.e. when $|m_{ki}^{(j)} - M_i| \lesssim p\cdot\sigma_{ki}$ ($p\cdot\sigma_{ki}$ being of the order of the proton mass), k=1, 2; p=1÷3 depending on desired value of a significance level /5/. Nevertheless, we can also perceive two long belts of the width of $p\cdot\sigma_{1i}$($p\cdot\sigma_{2i}$) along the m_2(m_1) axis for each sort of particles. If we compare both of mass distributions (see Figure 2, upper histogram, and Figure 3), we shall find that just these belts determine lower limits of occurrence frequancy for particles heavier than deuterons if an analysis is carried out, for instance, using an univariate approach only. Accordingly, the inequality (1) treated as a selection criterion of particles should be complete by adequate additional condition:

$$\begin{aligned} &(m_{1i}^{(j)} > p\cdot\sigma_{1i} \mid (m_{2i}^{(j)} - M_i) \le p\cdot\sigma_{2i}) \\ \text{or} \quad &(m_{2i}^{(j)} > p\cdot\sigma_{2i} \mid (m_{1i}^{(j)} - M_i) \le p\cdot\sigma_{1i}). \end{aligned} \quad (2)$$

Now we can apply this complex criterion, i.e. ((1) or (2)) as a selection rule to single out from a sample, in particular such particles whose production probability is very small in comparison with others. The result of such a selection is shown in Figure 2 (middle and lower histograms).

Finally, we have to estimate an efficiency of the method. For this purpose one can calculate from a sample of experimental data an admissible minimal value of the ratio $y = \sigma(Z_n, M_n) / \sum_{m\ne n}\sigma(Z_m, M_m)$ for particles of the n-th sort being of interest and producing very small signal. Qualitatively this can be done by means of the inequality:

$$y \ge \sum_{m\ne n} \alpha_m \cdot \beta_{nm}. \quad (3)$$

Here α_m is the significance level associated with a mass distribution of particles of the n-th sort, and β_{nm} is the probability of a Type II error, i.e. when a particle of the m-th sort is taken as a particle of the n-th sort. Numerical values of these probabilities (α_m and β_{nm}) can be estimated directly from the (m_1,m_2) scatter plot as shown in Figure 3. In the case under consideration (see Figure 3) we can get for y, using our complex criterion ((1) or (2)) at p=3, the value significantly smaller than 10^{-6}. Therefore the mass distribution for tritons (middle histogram in Figure 2) is practically without background. The same concerns the lower histogram (Figure 2), too where mass distributions for particles with Z=2 are displayed.

4. CONCLUSION

The particular case taken from experimental high energy physics and described in the paper proves that the multivariate approach to analysing data, whenever possible, may give an appreciable advantage over the univariate one. This remains true even if measured values taken as random variables are correlated to a certain degree (see, for example, /6/). Moreover, often it is not necessary to use more complicated or sophisticated statistics as selection criterions (tests) whose power may turn out remarkable smaller and their application may cause in practice even some difficulties (as, for instance, ω^2 statistic in /5,6/).

Some numerical results used in this work have been published earlier /4-6/.

ACKNOWLEDGEMENTS

I wish to express my appreciation to Professor M.G.Meshcheryakov for his permanent interest and encouragement.

REFERENCES

1. L.M.Anderson, Jr., Ph.D. thesis, Lawrence Berkeley Laboratory Report LBL-6769, 1977; L.M.Anderson et al. Phys.Rev.C, 1983, v.28, N.3, p.1224.
2. V.G.Ableev et al. JINR, 13-10568, Dubna, 1977.
3. L.S.Azghirey et al. JINR, D2-82-568, Dubna, 1982, p.83.
4. B.Słowiński et al. JINR, P10-86-831, Dubna, 1986.
5. B.Słowiński et al. JINR, P1-87-51, Dubna, 1987.
6. B.Słowiński et al. JINR, P10-86-832, Dubna, 1986.

Received by Publishing Department
on May 26, 1987.

Словинский Б. E11-87-367

Простой многомерный статистический подход к выделению слабых сигналов

Описаны два примера применения многомерного анализа данных. Эти примеры довольно типичны для экспериментальной физики высоких энергий. Они иллюстрируют преимущество даже простого многомерного статистического подхода к анализу численных результатов, если такой подход возможен, по сравнению с одномерным подходом. Показано также, что такой подход может быть построен в виде набора простых и быстрых процедур, пригодных для работы экспериментальной установки на линии с вычислительной машиной.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Препринт Объединенного института ядерных исследований. Дубна 1987

Słowiński B. E11-87-367

Simple Statistical Multivariate Approach to Weak Signals Extraction

In the paper two examples of application of the multivariate data analysis are described. These examples are typical enough for experimental high energy physics and illustrate an advantage of even simple multivariate approach to analysing numerical results, whenever possible, over the univariate one. It is pointed out too that such approach may be constructed as a set of simple fast procedures suitable for using when an experimental device operate on-line with a computer.

The investigation has been performed at the Laboratory of Computing Techniques and Automation, JINR.

Preprint of the Joint Institute for Nuclear Research. Dubna 1987