Physics of Atomic Nuclei, Vol. 68, No. 1, 2005, pp. 50–59. Translated from Yadernaya Fizika, Vol. 68, No. 1, 2005, pp. 51–60.

ELEMENTARY PARTICLES AND FIELDS
Experiment

Application of a Multivariate Statistical Technique to Interpreting Data from Multichannel Equipment for the Example of the KLEM Spectrometer

D. M. Podorozhnyi, E. B. Postnikov*, L. G. Sveshnikova, and A. N. Turundaevsky

Institute of Nuclear Physics,
Moscow State University, Vorob'evy gory, Moscow, 119899 Russia

Received November 25, 2003; in final form, February 20, 2004

Abstract—A multivariate statistical procedure for solving problems of estimating physical parameters on the basis of data from measurements with multichannel equipment is described. Within the multivariate procedure, an algorithm is constructed for estimating the energy of primary cosmic rays and the exponent in their power-law spectrum. They are investigated by using the KLEM spectrometer (NUCLEON project) as a specific example of measuring equipment. The results of computer experiments simulating the operation of the multivariate procedure for this equipment are given, the proposed approach being compared in these experiments with the one-parameter approach presently used in data processing.

INTRODUCTION

Highly precise measurements of energy are required in order to solve many important problems in cosmic-ray physics—for example, in order to localize the break in the power-law spectrum of primary cosmic rays. The smaller the error in measuring energy, the higher the probability of correctly interpreting these measurements and the higher the accuracy in determining the parameters of the break in the spectrum and other features of the energy spectrum in a given range. Requirements for the accuracy in measuring energy become especially stringent in the case of low statistics, since, among all statistical factors, it is the volume of statistics that has the strongest effect on the magnitude of errors in determining the parameters of the break [1].

With the aid of modern measuring equipment, one can obtain vast amounts of digitized information. For example, a strip silicon detector that is used in the KLEM spectrometer (NUCLEON project [2, 3]) to record the angular distribution of secondary particles makes it possible to measure simultaneously pulse heights in a few hundred channels, each of the channels carrying information about the primary-particle energy. In the present study, we propose, for the example of a computer model of the KLEM spectrometer, a multivariate procedure for processing data obtained by recording cosmic rays (it should be emphasized, however, that problems of this type admit a similar solution for any multichannel detector or any multiparameter measuring equipment).

The NUCLEON project is aimed at developing recording equipment that is intended for studying cosmic rays (protons and nuclei) over a broad energy range and which would be characterized by a relatively low weight and a high sensitivity. The KLEM measuring procedure essentially consists in determining the primary-particle energy from the lateral density distribution $\rho(x, y)$ of the flux of secondary particles produced in a thin target (first inelastic-interaction event) and bred in an ultrathin push-out device [4]. Two strip-detector matrices orthogonal to each other, the signal N_i from each of the strip detectors being proportional to the ionization loss in the ith strip, are used to measure $\rho(x, y)$. We will refer to the signal N_i or to any other data of multichannel measuring equipment as measured variables and to physical quantities (for example, primary-particle energy) to be determined on the basis of these measurements as estimated parameters.

We will consider two types of problems that can be solved optimally—that is, to the highest possible precision within a broad class of algorithms. These are problems of deriving estimates on the basis of multivariate data from multichannel equipment—first, one or a few physical quantities not measurable directly [5] (for example, primary energy, charge, etc.) and, second, the exponent of the power-law primary spectrum.

*e-mail: postn@rbcmail.ru

1063-7788/05/6801-0050$26.00

1. DETERMINATION OF PRIMARY ENERGY IN EACH INDIVIDUAL EVENT

The simplest multivariate method—this is the method of obtaining, for a random vector, a linear estimate that corresponds to the best (least) mean-square deviation—makes it possible to derive, for the problem of estimating, on the basis of data from measurements with multichannel equipment, one or a few physical parameters not measurable directly, a solution that would be more precise than that provided by any other linear algorithm for their determination [6]. For the KLEM spectrometer, this statement implies that, for any choice of coefficients of the measured variables N_i in an empirical or a speculated formula for estimating the primary energy,

$$E = \sum_i C_i N_i + C_0,$$

the estimated value of E would not be better than that which is obtained by applying the multivariate procedure.

Despite the linearity of the method in question, its multivariate character by far compensates for this restriction: although any of the physical quantities measured experimentally is only taken into account within a linear dependence, numerous relations between the measured variables and the estimated parameter, as well as the interplay of the measured variables themselves, are included in the procedure in the best possible way. In practice, this algorithm therefore works much better than any "simplified" procedure of data treatment via replacing all variables measured with the aid of expensive equipment by one variable representing their combination, whereupon one constructs a nonlinear dependence of an unknown quantity on this variable. By way of example, we indicate that, within the method developed previously by our group [4] for determining primary energy by means of the KLEM spectrometer, one replaces a few thousand variables N_i by only one variable

$$S = \sum_{i=1}^{m} \{\ln^2(2r_i/H)\} N_i, \qquad (1)$$

where r_i is the distance between the ith strip, which recorded the signal N_i, and the axis of a shower of secondary particles and H is the distance between the strip-detector plane and the interaction point. Moreover, there are various methods for taking nonlinearities into account even within multivariate strategies in the case where the importance of these nonlinearities is suggested by physical considerations.

In order to realize this method, the energy E is treated as a random variable, while all of the m measured variables (for example, signals from the detector strips) are treated as the coordinates of an m-dimensional random vector (it is denoted here by $\boldsymbol{\xi}$). After that, a linear estimate of the quantity E is formed on the basis of the entire body of information available from measurements; that is,

$$E_{\text{est}} = \sum_{i=1}^{m} b_i \xi_i + c, \qquad (2)$$

where the constant coefficients b_i and c are chosen in such a way as to minimize the mean-square deviation of the estimate of E from its true value (mean-square error):

$$\mathrm{M}(E_{\text{est}} - E)^2 = \left(\sum_{i=1}^{m} b_i \xi_i + c - E \right)^2 \sim \min \quad (3)$$

$$\{\text{in relation to any other linear estimate of } E\}.$$

Here, M symbolizes the expectation value.

The sought values of the coefficients appearing in the formula for estimating E are given by the theorem quoted in [6]; that is,

$$b_i = (\mathbf{S}_{E\xi} \cdot \mathbf{S}_{\xi}^{-1})_i, \quad c = \mathrm{M}E - \mathbf{S}_{E\xi} \cdot \mathbf{S}_{\xi}^{-1} \mathrm{M}\boldsymbol{\xi}, \quad (4)$$

where $\mathbf{S}_{\xi}$ is the autocovariance matrix for the random vector $\boldsymbol{\xi}$, $\mathbf{S}_{E\xi}$ is the mutual covariance matrix for E and $\boldsymbol{\xi}$, and the index i after a parenthesis labels the ith coordinate of a vector. Instead of unknown covariance matrices and expectation-value vectors, we use their unbiased estimates obtained on the basis of data from a learning sample (that is, a sample characterized by a rather large volume and specially simulated for estimating unknown coefficients), for example,

$$\mathrm{M}\xi \approx \langle \xi \rangle = \frac{1}{n_t} \sum_{i=1}^{n_t} \boldsymbol{\xi}_i, \qquad (5)$$

$$\mathbf{S}_{E\xi} = \frac{1}{n_t - 1} \sum_{i=1}^{n_t} (E_i - \langle E \rangle)(\boldsymbol{\xi}_i - \langle \boldsymbol{\xi} \rangle)^{\mathrm{T}},$$

where n_t is the volume of the learning sample; $\boldsymbol{\xi}_i$ and E_i are the ith realizations of the vector $\boldsymbol{\xi}$ and the energy E, respectively; angular brackets denote averaging; and T denotes transposition.

In general, the algorithm used to estimate energy involves the following steps:

(i) The response of the device to the passage of a beam of primary particles through the measuring equipment is simulated, their energy spectrum being preset; in other words, there occurs the formation of a learning sample.

(ii) The algorithm of estimation by formulas (2), (4), and (5) is formulated.

In addition, one can incorporate, into the procedure being developed, one or a few extra parameters that would describe our a priori ideas of the character

of the statistical relationship (more precisely, of the nonlinearity present in this relationship, since all of the linear correlations are estimated automatically) between measured and (or) estimated variables. The parameters themselves can be chosen on the basis of physical considerations, while their optimum values are fixed by using the results of a numerical (computer) experiment that simulates the operation of the procedure being developed. It follows that the formation of yet another random sample (a test one) is necessary, and this is the next step.

(iii) The test sample is formed by simulating the operation of the measuring device, and a computer testing of the procedure of estimation is performed on the basis of this new sample. After that, the error in the estimation on the basis of (3) is calculated, and the optimum values of all unknown parameters of the method are determined as those that minimize the error of the estimation.

As to the form of the energy spectrum of the learning and test samples, it must be determined by the special features of a concrete applied problem to be solved by means of the above algorithm for estimating the primary-particle energy. In order to improve the accuracy of the estimation, it is necessary to include a greater amount of various a priori information about the physical process being studied. Yet another important comment is in order. For a criterion that the procedure being developed must satisfy, one can take not only the condition in (3), which requires that the absolute error in estimating energy, $\mathrm{M}(E_{\mathrm{est}}-E)^2$, be minimized. The algorithm in question can be modified in such a way that it would minimize the dimensionless relative error $\mathrm{M}((E_{\mathrm{est}}-E)/E)^2$, which has a clearer meaning. The condition in (3) will then assume the form

$$\mathrm{M}((E_{\mathrm{est}}-E)/E)^2 \tag{6}$$
$$=\left(\left(\sum_{i=1}^{m} b_i\xi_i + c - E\right)/E\right)^2 \sim \min$$
$$\{\text{in relation to any other linear estimate of } E\}.$$

After some simple algebra, this problem reduces to the preceding one. The ingredients of the algorithm described in items (i)–(iii) and used to determine the primary-particle energy undergo no changes, with the exception of the formula for determining, on the basis of a simulated learning sample, the coefficients b_i and c in expression (2) for E_{est}; that is,

$$b_i = (\boldsymbol{\Sigma}_{\tilde{E}\tilde{\boldsymbol{\xi}}}\cdot\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\xi}}}^{-1})_i, \tag{7}$$
$$c = \{1/\mathrm{M}(1/E^2)\}\{\mathrm{M}(1/E) - \boldsymbol{\Sigma}_{\tilde{E}\tilde{\boldsymbol{\xi}}}\cdot\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\xi}}}^{-1}\mathrm{M}(\boldsymbol{\xi}/E^2)\},$$

where $\tilde{\boldsymbol{\xi}} = \{1/E\}\{\boldsymbol{\xi} - \mathrm{M}(\boldsymbol{\xi}/E^2)/\mathrm{M}(1/E^2)\}$; $\tilde{E} = 1-(1/E)\mathrm{M}(1/E)/\mathrm{M}(1/E^2)$; and $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\xi}}}$ and $\boldsymbol{\Sigma}_{\tilde{E}\tilde{\boldsymbol{\xi}}}$ are correlation (that is, noncentered) matrices, whose sample estimates are obtained in a way similar to that in (5):

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\xi}}} = \frac{1}{n_t}\sum_{i=1}^{n_t}\tilde{\boldsymbol{\xi}}_i\cdot\tilde{\boldsymbol{\xi}}_i^{\mathrm{T}}, \quad \boldsymbol{\Sigma}_{\tilde{E}\tilde{\boldsymbol{\xi}}} = \frac{1}{n_t}\sum_{i=1}^{n_t}\tilde{E}_i\tilde{\boldsymbol{\xi}}_i^{\mathrm{T}}. \tag{8}$$

In contrast to what we have in (2) and (4), the estimate of energy, E_{est}, is no longer unbiased upon such a modification; that is, $\mathrm{M}E_{\mathrm{est}} \neq \mathrm{M}E$.

Finally, we would like to dwell at some length on the parameters that make it possible to take into account, within the chosen procedure, the nonlinearity of the physical processes being considered. First, we note that, even in the course of computer experiments that relied on a one-dimensional algorithm for estimating the primary-particle energy and which employed the artificial variable S (1), it was found that the tightest correlation is observed between E and N_i^a, where $a \approx 1.2$–1.4. This circumstance, as well as the case where the tightest correlation would take place between an unknown parameter (E) and any arbitrarily complicated known function of measured variables, can readily be taken into account within the multivariate algorithm for estimation as well. In order to include this a priori information, it is sufficient to modify appropriately, from the outset, the input database and to employ, in the following, data on N_i^a rather than on the recorded signals N_i themselves.

Yet another factor that enables one to take efficiently into account physical processes underlying the operation of the measuring equipment is inherent in the computational procedure of the multivariate method for estimation [7]. The point is that we realized the algorithm of pseudoinversion of the correlation matrix [8] with the aid of only a few maximal singular quantities whose number is determined in a computer experiment as that which minimizes the error in estimating energy. This algorithm implements some kind of "filtration" of small-scale, insignificant, and spurious interrelations stemming from insufficiently vast statistics and concurrently removes difficulties associated with addressing ill-posed problems.

As applied to the KLEM spectrometer, the algorithm for estimating energy is the following:

(i) The form of the cosmic-ray energy spectrum that is proposed to be recorded in a simulated or an actual experiment is chosen (for example, a power-law spectrum or a few monochromatic beams of fixed energy). The form of the learning and the test sample is chosen accordingly for a subsequent accumulation of computer statistics. The type of error—the absolute

error, as in (3), or the relative error, as in (6)—is chosen.

(ii) The beam of primary particles belonging to the sort in which we are interested and having the preset form of the energy spectrum (see the preceding item) is simulated at the input of the computer model of the detector. Preset values of the primary energy E and the measured values of the signal N_i are successively recorded in a file for each of the simulated events that involve the passage of beam particles through the device. The learning sample E, N_{i1}, E, N_{i2}, ..., E, N_{in} is formed, and all N_i are transformed into N_i^a, where the constant a is taken to be unknown for the time being.

(iii) The unknown constants of the algorithm are evaluated by formulas (4) and (5) or (7) and (8).

(iv) The testing sample is formed in a way similar to that described in item (i) of the algorithm, the energy of each particle from this sample is estimated, and the error in (3) or in (6) is calculated by means of averaging over the entire sample. The optimum value at which the error is minimal is determined for the parameter a. Formula (2) for estimating the primary energy has now been fully specified, since the values of all constants appearing in it have been determined.

(v) A test beam having the structure, spectral shape, and intensity in which were are interested is transferred to the input of the computer model of the measuring equipment (and, in the future, to the input of the actual device); the energy of each particle in the beam is estimated; and the error in these quantities is calculated.

We will now present the results obtained by applying the above procedure to solving two problems within one general problem of reconstructing the energy of primary particles. The first of these is that of determining the energy of each particle from a beam having a power-law energy spectrum, while the second is that of determining the energies of particles from a few beams of monochromatic energy between $E = 10^{11}$ and $E = 10^{15}$ eV.

Only protons incident orthogonally to the measuring-equipment plane were simulated in all of the cases considered here. This simulation was based on the GEANT 3.21 package [9]. High-energy interactions of hadrons were described with the aid of the QGSJET generator [10], while their low-energy interactions (up to 50 GeV) were treated by using the FLUKA generator [9]. The applicability of these models to describing hadron interactions was confirmed by a comparison with experimental data [10, 11].

Within the first problem, it is assumed that we know the shape of the actual energy spectrum. This is a power-law function, but it is not necessary that its exponent be known to a high precision. In order to take into account this a priori information, a random sample for learning our procedure must be taken precisely from a power-law distribution characterized by the presumed exponent value γ. For a criterion, we took the relative error. According to the results of the simulation, the mean-square error in estimating energy was 49%. The one-dimensional method that employs the variable S yields 56%. These values receive overwhelming contributions from low-energy events, since they result from averaging over a steeply descending power-law spectrum.

In the second problem, where monochromatic beams of energy ranging between 10^{11} and 10^{15} eV form the test sample, we are equally interested in energy values over the entire range on a logarithmic scale, from $E = 10^{11}$ to $E = 10^{15}$ eV; therefore, a random sample from an energy spectrum such that the logarithm of energy is uniformly distributed over the entire range that we chose must be taken to be learning. Such a sample was formed by about 500 events over the entire energy range covering five orders of magnitude. The volume of each of six test samples monochromatic in energy ranged between 100 and 500 events.

The results of estimating energy are given in Table 1 for several values of the parameter a, which characterizes nonlinearity. In order to compare these results with those that emerge from the application of the already existing procedure used within the KLEM–NUCLEON project to estimate energy, similar errors were calculated for the same test samples by means of the algorithm based on the single variable S in (1). We note that, in contrast to what was done previously in [4], we did not perform any low-energy truncation in calculating these errors—we took into account the entire body of statistics generated for test samples. Moreover, it is of importance that, as a matter of fact, the errors of the earlier procedure were calculated for the same data as those that were used in the algorithm itself to construct the calibration curve (thus, it was an a priori known energy that was subjected to reconstruction). This means that, within the earlier algorithm, it would be natural to expect even a poorer accuracy of reconstruction for different samples.

The last column of Table 1 gives the results obtained previously in [4] from computer experiments aimed at estimating the primary-particle energy with the KLEM spectrometer, where the error was calculated by using incomplete statistics, its part at the lowest energies being eliminated. Albeit being incorrect from the mathematical point of view, this procedure did not lead to loss of information significant for the ensuing data treatment, since the reconstructed energy values were further used directly to construct

Table 1. Relative error in reconstructing energy (in %)

E, eV	Multivariate method				One-dimensional method employing the parameter S	
	$a = 1.2$	$a = 1.3$	$a = 1.4$	$a = 1.5$	full statistics	result obtained in [4] without the low-energy part of statistics
10^{11}	46	43	37	33	92	72
10^{12}	58	59	62	64	103	69
10^{13}	61	61	62	64	101	61
10^{14}	60	62	63	65	95	55
10^{15}	63	66	69	73	83	56

the primary spectrum by means of histograms. The region of low energies is of no interest from this point of view, whereas the tail in the region of underestimated values of E_{est} makes a significant contribution to the total error, as is suggested by a comparison of the data in the last two columns of Table 1. The algorithm that will be proposed in the present study for reconstructing the exponent in the power-law spectrum employs only an estimate of $\langle \ln E \rangle$ rather than estimates of energy.

2. DETERMINATION OF THE EXPONENT IN THE PRIMARY POWER-LAW SPECTRUM

In order to reconstruct the exponent γ in the energy distribution of cosmic rays,

$$p(E) = \frac{\gamma - 1}{E_0} \left(\frac{E}{E_0} \right)^{-\gamma} \qquad (9)$$

(E_0 is the left boundary of the spectrum), and the shape of the spectrum, we previously used the traditional procedure for plotting histograms on the basis of reconstructed energy values [12, 13]. This procedure involves a large error, which is difficult to estimate, but it has long since become a conventional tool in these realms. Since the method proposed in Section 1 for estimating energy leads to unbiased results and involves a minimum mean-square error, it is quite natural to expect that even a direct application of the traditional algorithm of reconstructing the spectrum at energy values found by the new method, which is not in use at the present time, would lead to a higher precision in reconstructing the spectrum.

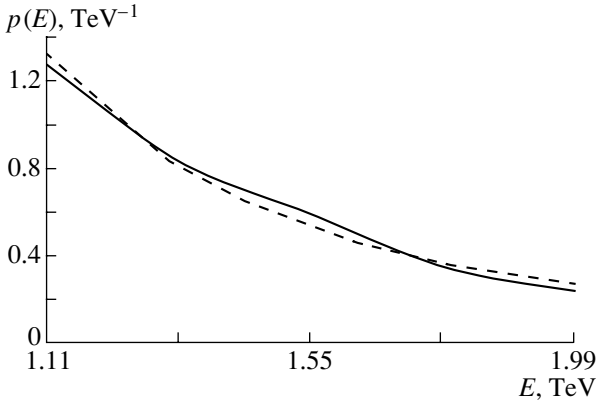

Fig. 1. Density of the primary energy distribution (solid curve, based on an analytic form) before and (dashed curve, constructed on the basis of a histogram) after the trigger.

One of the most serious difficulties in reconstructing the spectrum is that, in simulating the operation of the KLEM spectrometer, one performs a "multistep" selection of events that the detector used would record. As a result, the shape of the primary spectrum is severely distorted, so that even a perfectly precise measurement of energies of particles recorded by the detector would give no way to reconstruct their spectrum at the input of the measuring equipment (Fig. 1). The selection criterion results in that the exponent γ calculated by the maximum-likelihood method for the primary spectrum having the lower boundary at $E_0 = 1$ TeV is underestimated to become $\gamma_{\text{est}} = 2.58$ (in the case of a precise measurement of the energies of all particles that passed a triggering selection of particles) instead of $\gamma_0 = 2.70$, whereas, for the same volume of the sample, the statistical uncertainty in estimating γ can be determined as

$$\sigma_{\gamma_{\text{est}}} \sim \sigma\{1/\ln(E/E_0)\} = 0.017$$

[on the basis of formula (10) below, which provides a realization of the method in question in the case of precise measurements].

2.1. Spectrum Unbounded from the Right

In order to estimate the exponent γ in the power-law distribution by the maximum-likelihood method (MLM), one can make use of the formula

$$\gamma_{\text{est}} = 1 + 1/(\langle \ln E \rangle - \ln E_0). \qquad (10)$$

The quantity $\langle \ln E \rangle$ can be found by averaging the logarithms of the measured energy values only if the energy is measured without errors or if the errors in determining $\ln E$ do not depend on energy, but, in either case, this is an idealization—otherwise, the distribution of measured energies is the convolution of the primary spectrum with a function that describes distortions introduced by the measuring device. In the problem at hand, an additional distortion of the spectral shape arises even at the preliminary stage of event selection by instrumental triggers. Therefore, we applied a procedure that immediately yields the most precise estimate of $\ln E$—namely, a linear estimate that is constructed for $\ln E$ treated as a random variable and which is the best in the sense of the mean-square deviation. In contrast to the method employing the parameter S, this method yields unbiased results, guaranteeing that the respective estimate of $\langle \ln E \rangle$ will not suffer from systematic under- or overestimations.

The algorithm used to estimate the exponent of the primary power-law spectrum is the following:

(i) The "preliminary" step consists in choosing a few values of $\gamma = \gamma_0$ from the interval in which we are interested. In all, we employed three values of γ_0 in our numerical experiments (this was sufficient to ensure a fairly high precision).

(ii) For each of the chosen values of γ_0, two random samples are taken from the power-law spectrum that has this exponent. The use of precisely a power-law distribution for learning the method involves taking into account additional a priori information. In the procedure implemented further to reconstruct the logarithm of the energy for each γ_0, one sample will be used as a learning one, while the other is taken to be a test one.

(iii) For each individual event of the test sample, $\ln E$ is reconstructed by formulas (2), (4), and (5), where the random variable E is replaced by $\ln E$. For a learning sample, we employ that which features fixed γ_0 (beginning with the first one), while, for a test sample, we successively take samples involving each of the three values of γ_0 (including that which corresponds to the learning sample).

(iv) For each of the three sets of $\ln E$ that were determined at the preceding step, a preliminary estimate of γ is found by formula (10). These will be "preliminary" estimates of γ, the true values being equal to the first, the second, and the third of the γ_0 values, respectively; the learning of the procedure was performed by using one (initially, the first) of these γ_0 values.

(v) The procedures of steps (iii) and (iv) are repeated by using, for a learning sample, the sample that involves, first, the second and, then, the third value of γ_0. Thus, we performed the procedure for reconstructing $\ln E$ three times for each of the three γ_0 values, thereby deriving nine sets of reconstructed values of $\ln E$ and the corresponding "preliminary" estimates of γ_0: three estimates for the first value of γ_0, three estimates for the second one, and three estimates for the third one. The three estimates of the same value of γ_0 differ in that different learning samples (successively, the samples involving the first, the second, and the third value of γ_0) were used to obtain them.

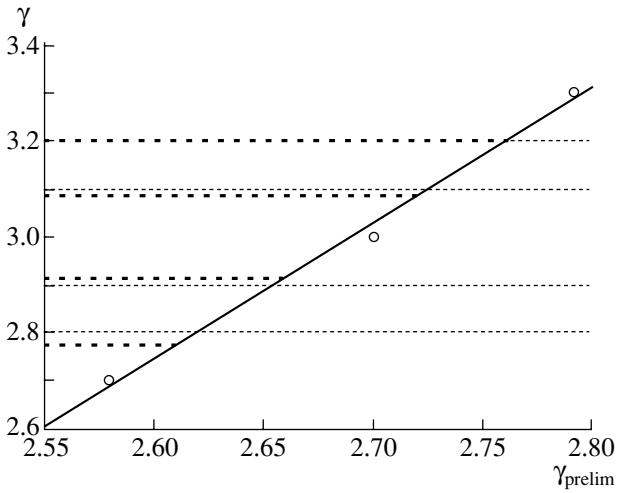

Fig. 2. Estimating γ on the basis of the linear dependence $\gamma(\gamma_{\text{prelim}})$: (open circles) "learning" points, (solid line) interpolation straight line corresponding to the least squares method, (thick dashed lines) ultimate estimates of the parameter γ, and (dotted lines) true values of the estimated γ.

(vi) An interpolation curve representing the dependence of the true value of γ on its "preliminary" estimate is constructed on the basis of the points found at the preceding step (in our case of three points, we use a linear function). For each of the three values of γ_0, we construct an individual interpolation dependence. For cases like that in which the estimated value is much greater or much less than the known one, we thereby obtain the possibility of comparing the quality of the developed procedure for different values of γ_0 preassigned for learning this procedure.

(vii) The "ultimate step" consists in finding the estimates of the exponent γ that are corrected with the aid of the three interpolation dependences constructed at the preceding step. An example of how the procedure outlined here is represented graphically is given in Fig. 2.

We have performed computer experiments aimed at estimating the exponent of a power-law proton spectrum (for a vertical incidence of the beam to the detector plane). In order to compare our multivariate

Table 2. Estimates of the exponent γ in the form $\langle\gamma_{\text{est}}\rangle \pm \Delta_\gamma$

True values of γ	$N = 100$, $\sigma_{\text{MLM}} = 0.18$				$N = 200$, $\sigma_{\text{MLM}} = 0.12$		$N = 300$, $\sigma_{\text{MLM}} = 0.10$		$N = 400$, $\sigma_{\text{MLM}} = 0.09$	
	One-dimensional method involving the parameter S	Multivariate method			One-dimensional method involving the parameter S	Multivariate method, $\gamma_0 = 2.7$	One-dimensional method involving the parameter S	Multivariate method, $\gamma_0 = 2.7$	One-dimensional method involving the parameter S	Multivariate method, $\gamma_0 = 2.7$
		$\gamma_0 = 2.7$	$\gamma_0 = 3.0$	$\gamma_0 = 3.3$						
2.8	2.81 ± **0.29**	2.81 ± **0.30**	2.81 ± **0.31**	2.81 ± **0.30**	2.80 ± **0.21**	2.80 ± **0.22**	2.81 ± **0.15**	2.79 ± **0.17**	2.78 ± **0.13**	2.79 ± **0.15**
2.85	2.90 ± **0.33**	2.83 ± **0.32**	2.85 ± **0.31**	2.85 ± **0.30**	2.88 ± **0.22**	2.84 ± **0.22**	2.84 ± **0.13**	2.83 ± **0.18**	2.86 ± **0.13**	2.84 ± **0.16**
2.9	2.90 ± **0.31**	2.89 ± **0.30**	2.91 ± **0.29**	2.91 ± **0.30**	2.92 ± **0.24**	2.89 ± **0.21**	2.88 ± **0.16**	2.89 ± **0.17**	2.87 ± **0.14**	2.89 ± **0.16**
3.1	3.19 ± **0.46**	3.14 ± **0.27**	3.16 ± **0.26**	3.13 ± **0.27**	3.11 ± **0.29**	3.14 ± **0.19**	3.13 ± **0.20**	3.13 ± **0.15**	3.15 ± **0.19**	3.13 ± **0.13**
3.15	3.24 ± **0.44**	3.17 ± **0.26**	3.17 ± **0.25**	3.17 ± **0.24**	3.22 ± **0.34**	3.18 ± **0.18**	3.20 ± **0.25**	3.17 ± **0.15**	3.16 ± **0.19**	3.18 ± **0.13**
3.2	3.20 ± **0.43**	3.16 ± **0.26**	3.16 ± **0.25**	3.17 ± **0.26**	3.13 ± **0.25**	3.15 ± **0.18**	3.09 ± **0.21**	3.16 ± **0.15**	3.13 ± **0.19**	3.15 ± **0.14**

procedure with that which employs one parameter, the exponent γ was estimated by the two methods as applied to the same simulated data—that is, by the algorithm that employs a multivariate statistical estimation of the logarithm of energy and by the method that reconstructs energy on the basis of the parameter S (1).

The learning of the multivariate method was performed by using three proton beams having a power-law energy spectrum whose exponent γ_0 takes the values of 2.7, 3.0, and 3.3. The exponent was estimated for beams characterized by a set of γ values in the range between 2.8 and 3.2. The number of events in each of the learning beams was quite large (a few thousand), but this imposed no constraints on the implementation of the tested procedure in practice, since the learning samples can be accumulated via a computer simulation rather than in an actual experiment. The test samples were taken to have a volume of 100 to 400 events—such numbers of primary protons can be recorded by the KLEM facility on board a cosmic vehicle.

The results of the estimation are given in Table 2. For the purposes of visualization, the data in the column corresponding to $N = 100$ and $\gamma_0 = 2.7$ are represented graphically in Fig. 3. Since the estimates of γ that were obtained for each of the three learning samples specified by the values of $\gamma_0 = 2.7$, 3.0, and 3.3 proved to be close to one another, only estimates at $\gamma_0 = 2.7$ are given in all parts of the table, with the exception of the first one.

On the basis of the data in Table 2, one can assess the strength of the effect that the volume of accumulated data has on the accuracy of estimation. As was indicated above, the statistical uncertainty of an estimate due exclusively to the finiteness of a sample (the energy is known precisely) can be obtained by using the maximum-likelihood-method formula (10). This uncertainty is "irremovable"; therefore, it is of paramount importance to get an idea of the order of its magnitude playing the role of the "limiting resolution" (which corresponds to the case of perfectly accurate measurements) of the procedure (or facility) for reconstructing the exponent γ. A graph that represents this "irremovable" uncertainty as a function of the volume of statistics, N, is given in Fig. 4 for an interval covering a few hundred events, which is of interest to us.

We note that, although each estimation of γ by formula (10) involves only data associated with N (from 100 to 400) events in Table 2, the estimates of γ are averaged over a few N-event samples from the entire body of available data in order to suppress random "outliers" and to verify an unbiased character of the results given by this procedure and the absence of a systematic bias. However, we will have modest statistics in an actual experiment, and this will prevent averaging results over a few samples. Therefore,

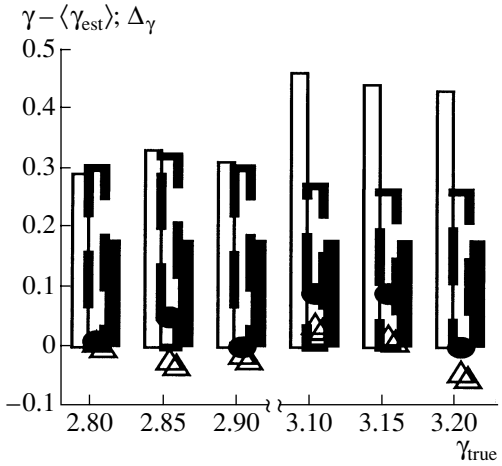

Fig. 3. Deviations from the mean value, $\gamma - \langle\gamma_{\rm est}\rangle$, and mean-square errors, Δ_γ, for estimates of the exponent γ (here, $N = 100$ and $\sigma_{\rm MLM} = 0.18$; for the multivariate method, $\gamma_0 = 2.7$): (thin-solid-line histogram) Δ_γ within the one-dimensional method, (thick-dashed-line histogram) Δ_γ within the multivariate method, (closed histogram) $\sigma_{\rm MLM}$, (closed circles) $(\gamma - \langle\gamma_{\rm est}\rangle)$ within the one-dimensional method, and (double triangles) $(\gamma - \langle\gamma_{\rm est}\rangle)$ within the multivariate method.

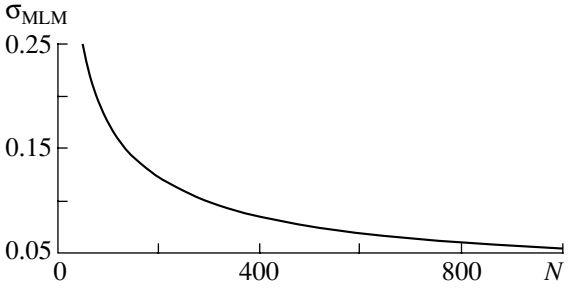

Fig. 4. Error in estimating γ by the maximum-likelihood method for a precisely known primary energy versus the volume of statistics.

the errors Δ_γ (printed in boldface type in Tables 2–4), which can be used to assess the degree of deviations from the averaged value that are expected in performing a single experiment for statistics involving N events, carry information of no less importance. The smaller the factor by which this error exceeds the "irremovable" error $\sigma_{\rm MLM}$, which is displayed in Table 2 and in Fig. 4, the higher the quality of estimation.

It should be noted that the above comparison of the two procedures, the one-dimensional and the multivariate one, involves some degree of arbitrariness, since no generally accepted algorithm for reconstructing the exponent of the power-law spectrum in processing data simulated for the KLEM equipment exists at the present time, and this was one of the reasons for developing a new universal algorithm. In our case, the values of E (or $\ln E$) that were obtained for each of the primary particles from simulated beams by the multivariate method and by the one-dimensional method employing the parameter S were merely subjected to identical treatment. It follows that, as a matter of fact, the same complicated algorithm of treatment was applied to the results of energy measurements by both procedures.

In order to render the conditions of our numerical experiment closer to those that will be prevalent in a live experiment, where the left boundary of the spectrum of recorded particles will not be known, the estimation of γ for an unknown left boundary of the spectrum was simulated in an independent run of the calculations. We took only those events that were selected according to the criterion $E_{\rm est} > E_0$, where E_0 is a known preset value (more rigorously, one does not determine the energy itself within the multivariate procedure of estimation; therefore, the selection criterion has the form $(\ln E)_{\rm est} > \ln E_0$ within this algorithm). The value for the left boundary of the spectrum was chosen, first, with allowance for the possibility of estimating it to a fairly high degree of precision and, second, with allowance for the volume of data that is necessary for the present purposes. On the basis of these considerations, we choose a few values of E_0 in the range between 2 and 4 TeV. Table 3 displays the results obtained by estimating γ for some values of E_0.

From a comparison of these results with the data in Table 2, it can be seen that the exponent of the spectrum whose left boundary is a priori unknown and is reconstructed on the basis of results of measurements performed with recording equipment can be estimated to a precision not poorer than that attained in estimating the exponent of the spectrum characterized by a fixed value of E_0.

2.2. Spectrum within the (E_1, E_2) Segment

In the case where one is interested in the value of the exponent γ only within some segment of the energy spectrum of primary cosmic rays, it is advisable to consider the spectrum in a form different from that in (9),

$$p(E) = \frac{\gamma - 1}{E_1^{1-\gamma} - E_2^{1-\gamma}} E^{-\gamma}, \quad \text{if} \quad E \in (E_1, E_2);$$

$$p(E) = 0, \quad \text{if} \quad E \notin (E_1, E_2).$$

This form of the spectrum is more complicated from the point of view of estimating γ, since, in this case, the maximum-likelihood method yields, instead of the direct formula (10), a nonlinear equation for γ,

$$\gamma = 1 + (E_1^{1-\gamma} - E_2^{1-\gamma})/(E_1^{1-\gamma}\{\langle\ln E\rangle - \ln E_1\} \tag{11}$$

Table 3. Estimates of the exponent γ in the form $\langle\gamma_{\text{est}}\rangle \pm \Delta_\gamma$ at a fixed left boundary E_0 of the spectrum for $\gamma_0 = 2.7$, 3.0, and 3.3 (here, $N = 100$ and $\sigma_{\text{MLM}} = 0.18$)

True values of γ	$E_0 = 2$ TeV			$E_0 = 2.5$ TeV			$E_0 = 3$ TeV		
	2.7	3.0	3.3	2.7	3.0	3.3	2.7	3.0	3.3
2.8	$2.84 \pm \mathbf{0.23}$	$2.82 \pm \mathbf{0.23}$	$2.83 \pm \mathbf{0.22}$	$2.85 \pm \mathbf{0.25}$	$2.86 \pm \mathbf{0.24}$	$2.86 \pm \mathbf{0.27}$	$2.83 \pm \mathbf{0.22}$	$2.83 \pm \mathbf{0.19}$	$2.80 \pm \mathbf{0.21}$
2.85	$2.83 \pm \mathbf{0.23}$	$2.83 \pm \mathbf{0.24}$	$2.80 \pm \mathbf{0.23}$	$2.81 \pm \mathbf{0.24}$	$2.79 \pm \mathbf{0.23}$	$2.76 \pm \mathbf{0.25}$	$2.80 \pm \mathbf{0.22}$	$2.77 \pm \mathbf{0.20}$	$2.80 \pm \mathbf{0.21}$
2.9	$2.91 \pm \mathbf{0.24}$	$2.91 \pm \mathbf{0.23}$	$2.88 \pm \mathbf{0.22}$	$2.90 \pm \mathbf{0.24}$	$2.91 \pm \mathbf{0.24}$	$2.92 \pm \mathbf{0.25}$	$2.93 \pm \mathbf{0.23}$	$2.93 \pm \mathbf{0.21}$	$2.93 \pm \mathbf{0.21}$
3.1	$3.11 \pm \mathbf{0.23}$	$3.10 \pm \mathbf{0.24}$	$3.08 \pm \mathbf{0.24}$	$3.14 \pm \mathbf{0.25}$	$3.10 \pm \mathbf{0.23}$	$3.11 \pm \mathbf{0.24}$	$2.97 \pm \mathbf{0.22}$	$2.97 \pm \mathbf{0.20}$	$3.02 \pm \mathbf{0.18}$
3.15	$3.25 \pm \mathbf{0.28}$	$3.24 \pm \mathbf{0.26}$	$3.22 \pm \mathbf{0.25}$	$3.27 \pm \mathbf{0.30}$	$3.26 \pm \mathbf{0.29}$	$3.33 \pm \mathbf{0.33}$	$3.26 \pm \mathbf{0.27}$	$3.23 \pm \mathbf{0.23}$	$3.27 \pm \mathbf{0.27}$
3.2	$3.19 \pm \mathbf{0.24}$	$3.17 \pm \mathbf{0.23}$	$3.16 \pm \mathbf{0.23}$	$3.26 \pm \mathbf{0.26}$	$3.30 \pm \mathbf{0.27}$	$3.30 \pm \mathbf{0.28}$	$3.26 \pm \mathbf{0.23}$	$3.20 \pm \mathbf{0.19}$	$3.20 \pm \mathbf{0.19}$

$$- E_2^{1-\gamma}\{\langle \ln E\rangle - \ln E_2\}).$$

The algorithm of estimation exactly reproduces that which was described above for the case of an unbounded spectrum, the only exception being that the maximum-likelihood method, which underlies both algorithms, is now realized through Eq. (11) rather than through formula (10).

As in the case of an unbounded spectrum, the vertical incidence of a proton beam to the detector plane was considered in computer experiments aimed at estimating the exponent γ within various energy ranges. All of the parameters of the simulation were identical to those in the preceding case. The values for both the left and the right boundary of the spectrum were not considered to be known and were reconstructed on the basis of simulated data, as is described in the preceding subsection.

For statistics including 100 events, Table 4 shows the results for an energy interval of width 2 TeV. It can be seen that, even for so small a volume of data, the exponent γ can be estimated within an energy range of small width by using the proposed procedure, albeit the uncertainty is somewhat greater than for an unbounded spectrum. This is because the irremovable error inherent in the maximum-likelihood method is greater in this case. This error now depends not only on the volume of statistics but also on the width of the energy interval. The relationship between the actual and the minimum possible error remains approximately identical to that in estimating the exponent γ of an unbounded spectrum.

Table 4. Estimates of the exponent γ in the form $\langle\gamma_{\text{est}}\rangle \pm \Delta_\gamma$ for the interval $2 < E < 4$ TeV (here, $N = 100$ and $\sigma_{\text{MLM}} = 0.29$)

True values of γ	One-dimensional method involving the parameter S	Multivariate method		
		$\gamma_0 = 2.7$	$\gamma_0 = 3.0$	$\gamma_0 = 3.3$
2.8	$2.94 \pm \mathbf{0.50}$	$2.78 \pm \mathbf{0.37}$	$2.72 \pm \mathbf{0.36}$	$2.70 \pm \mathbf{0.35}$
2.85	$2.93 \pm \mathbf{0.45}$	$2.85 \pm \mathbf{0.39}$	$2.94 \pm \mathbf{0.37}$	$2.77 \pm \mathbf{0.36}$
2.9	$2.86 \pm \mathbf{0.46}$	$2.82 \pm \mathbf{0.41}$	$2.87 \pm \mathbf{0.37}$	$2.74 \pm \mathbf{0.39}$
3.1	$2.94 \pm \mathbf{0.48}$	$3.18 \pm \mathbf{0.40}$	$3.14 \pm \mathbf{0.36}$	$3.04 \pm \mathbf{0.37}$
3.15	$3.10 \pm \mathbf{0.48}$	$3.19 \pm \mathbf{0.41}$	$3.15 \pm \mathbf{0.36}$	$3.07 \pm \mathbf{0.38}$
3.2	$2.95 \pm \mathbf{0.52}$	$3.18 \pm \mathbf{0.40}$	$3.11 \pm \mathbf{0.37}$	$3.00 \pm \mathbf{0.41}$

CONCLUSION

Multivariate procedures for processing the results of measurements with multichannel equipment have been developed, implemented, and tested in computer experiments. These procedures, which are optimal within a broad class of algorithms in the sense that they are characterized by the highest sensitivity, are intended for estimating (i) physical parameters inaccessible to direct measurements (such as the primary energy and other features of the primary particle) and (ii) the exponent of the primary spectrum of cosmic rays.

The multivariate procedures for estimation have been studied in computer experiments employing a mathematical model for the KLEM measuring equipment from the NUCLEON project. The following conclusions have been drawn from the results of these experiments:

(a) In estimating the primary-particle energy, the multivariate procedure yields a much smaller error (by a factor of 1.5) in relation to the one-dimensional algorithm used previously by our group.

(b) In estimating the exponent of the spectrum of primary cosmic rays, the multivariate procedure works at least no poorer than the algorithm based on a one-dimensional estimation of the energy of each

individual event. At the same time, the new procedure in question, in contrast to algorithms that are aimed at determining the exponent of the spectrum from histograms on the basis of a one-dimensional estimation of energy and which were previously applied in the KLEM–NUCLEON project and in other investigations, is optimal in a rigorous mathematical sense, is universal, and makes it possible to employ codes of a single type in processing multiparameter data of any kind.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 03-02-16660).

REFERENCES

1. L. W. Howell, Nucl. Instrum. Methods Phys. Res. A **480**, 741 (2001).
2. G. L. Bashindzhagyan *et al.*, Preprint No. 99-13/571, NIIYaF MGU (Inst. Nucl. Phys., Moscow State Univ., Moscow, 1999).
3. J. Adams, G. L. Bashindzhagyan, P. G. Bashindzhagyan, *et al.*, Izv. Akad. Nauk, Ser. Fiz. **65**, 430 (2001).
4. N. A. Korotkova, D. M. Podorozhnyi, E. B. Postnikov, *et al.*, Yad. Fiz. **65**, 884 (2002) [Phys. At. Nucl. **65**, 852 (2002)].
5. E. B. Postnikov, G. L. Bashindzhagyan, N. A. Korotkova, *et al.*, Izv. Acad. Nauk, Ser. Fiz. **66**, 1634 (2002).
6. Yu. P. Pyt'ev, *Methods for Analyzing and Interpreting Experimental Results* (Mosk. Gos. Univ., Moscow, 1990), pp. 15, 16 [in Russian].
7. E. B. Postnikov, Candidate's Dissertation in Mathematics and Physics (Mosk. Gos. Univ., Moscow, 1999).
8. Yu. P. Pyt'ev, *Mathematical Methods for Interpreting Experiments* (Vysshaya Shkola, Moscow, 1989) [in Russian].
9. GEANT User's Guide, CERN DD/EE/83/1 (Geneva, 1983).
10. N. N. Kalmykov *et al.*, Preprint No. 98-36/537, NIIYaF MGU (Inst. Nucl. Phys., Moscow State Univ., Moscow, 1998).
11. I. D. Rapoport, A. N. Turundaevsky, and V. Ya. Shestoperov, Yad. Fiz. **65**, 176 (2002) [Phys. At. Nucl. **65**, 170 (2002)].
12. A. V. Apanasenko, V. A. Sukhadolskaya, V. A. Derbina, *et al.*, Astropart. Phys. **16**, 13 (2001).
13. I. P. Ivanenko *et al.*, in *Proceedings of the 23rd ICRC, Calgary, 1993* (*Contributed Papers*), Vol. 2, p. 17.

Translated by A. Isaakyan