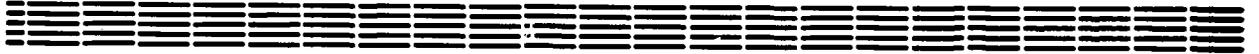


ԷԲԻ

Preprint YERPHI-1067(30)-88

ԵՐԵՎԱՆԻ ՖԻԶԻԿԱՅԻ ԻՆՍՏԻՏՈՒՏ
ЕРЕВАНСКИЙ ФИЗИЧЕСКИЙ ИНСТИТУТ
YEREVAN PHYSICS INSTITUTE



A.A.CHILINGARYAN, S.Kh.HARUTUNYAN

ON THE POSSIBILITY OF MULTIDIMENSIONAL
KINEMATIC INFORMATION ANALYSIS BY MEANS
OF NEAREST NEIGHBOUR DIMENSIONALITY
ESTIMATION

ЦНИИатоминформ
ЕРЕВАН—1988

Ո.Ք. ՀԱՐՈՒՓՅՈՒՆՅԱՆ, Ա.Ա. ՉԻԼԻՆԳԱՐՅԱՆ

ԲԱՅՄԱՆՈՒՄ ԿԻՆԵՄԱՏԻԿԱԿԱՆ ԻՆՓՈՐՄԱՅԻԱՆ ՉԱՓՈՂԱՍԿՈՒԹՅԱՆ Կ-ՄՀ-Ի
ԳՆԱՀԱՏՄԱՆ ՄԻՋՈՅՈՒՎ ԿԵՐԼՈՒԾԵԼՈՒ ԶՆԱՐԱՎՈՐՈՒԹՅԱՆ ՄԱՍԻՆ

Ներկայացված է բարձր էներգիաների Ֆիզիկայի բազմամասնիկ վերջնական վիճակների վերլուծության նոր եղանակ: Այն կարող է հանձնարարվել վիճակների նախնական վերլուծության համար և կարող է ցուցումներ տալ զանազան կապված վիճակների գոյություն մասին: Այս եղանակը հնարավորություն է ընձեռում իշականացնել սվյալների երկչափ արտապատկերում, ինչը առանձնապես կարևոր է սվյալների անմիջական մշակման պայմաններում: Հեղինակներն օգտագործում են կոոլյացիոն չափողականության որոշման նոր ալգորիթմներ, որոնք հիմնված են հավանականության բազմաչափ լստության Կ-ՄՀ-ի գնահատականի վրա: Այս ալգորիթմները ալելի հարմար են և ճշգրիտ ի համեմատություն նախկինում առաջարկվածների: Դա ձեռք է բերվում ի հաշիվ բնական ստեղծված և կոոլյացիոն ինտեգրալի բաշխման առույթի օգտագործման: Ալգորիթմների արդյունավետությունը ստուգվել է հաղորդման մոդելները խաղարկման հաշվողական փորձերի շարքում:

Երևանի Ֆիզիկայի ինստիտուտ

Երևան 1988



С.Х. АРУТЮНЯН, А.А. ЧИЛИНГАРЯН

О ВОЗМОЖНОСТИ АНАЛИЗА МНОГОМЕРНОЙ КИНЕМАТИЧЕСКОЙ
ИНФОРМАЦИИ С ПОМОЩЬЮ КБС ОЦЕНОК РАЗМЕРНОСТИ

В работе описывается новый метод анализа множественных конечных состояний в физике высоких энергий. Метод не претендует на полное выявление динамики реакции и определение масс и ширины резонансов. Он может быть рекомендован для предварительного анализа событий с большой множественностью и способен дать указание на существование различных связанных состояний. Метод позволяет производить двухмерное отображение многомерных данных, что особенно важно при диалоговом режиме обработки данных. В работе использованы новые алгоритмы определения корреляционной размерности, на основе КБС оценок многомерной плотности вероятности. Эти алгоритмы более удобны и точны в сравнении с ранее предложенными, из-за введения естественного масштаба и учета функции распределения корреляционного интеграла. Работоспособность алгоритмов проверена в серии вычислительных экспериментов с монте-карло реализациями процесса адронорождения.

Ереванский физический институт

Ереван 1988

A.A.CHILINGARYAN, S.Kh.HARUTUNYAN

ON THE POSSIBILITY OF MULTIDIMENSIONAL
KINEMATIC INFORMATION ANALYSIS BY MEANS
OF NEAREST NEIGHBOUR DIMENSIONALITY ESTIMATION

A new method of analysis of multiple final states in high energy physics is described. It can be recommended for preliminary analysis of events with high multiplicity and is able to make indication at the existence of various bound states. The method allows to perform two-dimensional mapping of multi-dimensional data, this being particularly important in the dialog mode of data handling. New algorithms are used for the determination of correlation dimensionality, based on KNN estimation of probability density. These algorithms are more suitable and precise as compared to earlier suggested ones, due to the introduced natural scale and the account of the distribution function of the correlation integral. The algorithms correctness is checked in a series of the Monte-Carlo hadroproduction simulations.

Yerevan Physics Institute

Yerevan 1988

Introduction.

Recently, a great success in the description of complex systems behaviour was achieved by using geometrical representations. The generalized dimensionality originally introduced by Reni [1] and applied by Grassberger and Procaccia for the analysis of chaotic behaviour [2] proved to be highly fruitful in various applications, beginning from the description of crystal growth [3] and up to the analysis of star cluster [4] and quark-gluon plasma [5] .

On the other hand, the development of Mandelbrot's ideas about the fractal character of the Nature [6] also gave rise to a new understanding of complexity in the physical experiment.

The kinematic information about high-multiplicity reactions sharply increases and simultaneously there enhance difficulties as to detect so far unknown mechanisms of production of a given final state [7] . Effective mass distributions do not

allow to make any definite conclusion.

All available information on the reaction is embedded in the values of all possible random quantities induced by this reaction and measured in experiment. Events are concentrated in relatively small regions of phase space.

Essential inhomogeneity and complexity of the events patterns in phase space just brought us to the idea to use a fractal approach for the analysis of multiple production. Fractal set in a wide sense is a set whose structure is related to dimensionality [8]. Fractal analysis proves to be useful every time when the systems behaviour is characterized by attractivity. That is, final states are grouped in some subspace called attractor, whose dimension is less than that of the initial phase space [9].

It should be mentioned that there exist many different definitions of dimensionality and specific schemes to calculate them for finite sets [10], most of which go back to the first generalization of the dimensionality notion by Hausdorff [11]. However for cases important to physical experiments, most of these definitions are equivalent; therefore we'll prefer the methods allowing to obtain adequate estimates for large dimensions of initial space.

Strict mathematical definitions of dimensions as well as references to the appropriate references can be found in [12].

Highly useful proved to be the approach worked out by Procaccia et al. and L.Young [13,14] in the recent years, which allowed to generalize some most popular definitions of dimensionality and create a numerical method of calculation. Note

that the aim of the fractal approach is not to equip us with a ready theory but to formulate empirical facts on a geometrical language [15] for a subsequent comprehensive analysis.

1. Correlation Dimension and Its Relation to KNN

Estimation of Probability Density.

Procaccia showed that there exists an infinite set of different dimensions characterizing an attractor:

$$D_q = \frac{1}{q-1} \lim_{\ell \rightarrow 0} \ln \sum_{i=1}^{M(\ell)} p_i^q / \ell^n \ell. \quad (1)$$

The d -dimensional initial space where an attractor arises is divided into $M(\ell)$ cubes (boxes, cells, bins....), and in each of them a probabilistic measure p_i is determined. A cube volume is ℓ^d , q is an arbitrary real number. One can readily show that at $q \rightarrow 0$ the generalized dimension coincides with self similarity dimension:

$$D_c = \frac{\ln(\Delta M_{K+1}/M_K)}{\ln(\ell_{K+1}/\ell_K)}, \quad (2)$$

where ΔM_K is a number of self-similar objects occurring at fragmentation of scale at a K -th step, self similarity dimension in turn is tightly connected with the Hausdorff dimension:

$$D_{q \rightarrow 0} = D_c = - \lim_{\ell \rightarrow 0} \lim_{N \rightarrow \infty} \ln M(\ell) / \ln \ell, \quad (3)$$

where N is number of points on the attractor.

Practically, dimension is determined as a slope of the straight line that connects $M(\ell)$ and ℓ in double-logarithmic scale. To do so, one, of course, should be given by a series $\{\ell_i\}, i=1, \dots, K, K \gg 3$ and calculate the relevant series of values of $\{M(\ell_i)\}$ - number of cells of the size ℓ_i wherein the points of the studied set had fallen. At $q \rightarrow 1$ the generalized dimension reduces to the information one:

$$\mathcal{D}_{q \rightarrow 1} = \mathcal{G} = \lim_{\ell \rightarrow 0} \lim_{N \rightarrow \infty} \sum_{i=1}^{M(\ell)} p_i \ln p_i / \ln \ell \quad (4)$$

Most important for the applied cases is a correlation dimension \mathcal{D} corresponding to the case:

$$\mathcal{D}_{q=2} = \mathcal{D} = \lim_{\ell \rightarrow 0} \lim_{N \rightarrow \infty} \ln \sum_{i=1}^{M(\ell)} p_i^2 / \ln \ell \quad (5)$$

The correlation dimension is significant, firstly, because it characterizes local structure of the attractor, and secondly, because, as will be seen further, it can readily be calculated for dimensions of the initial space $d \gg 2$. The algorithm of direct counting of cells is rather tedious and is applicable only for the case $d \leq 2$. Clearly, at $d=10$ and fragmentation of each axis by 10, already at the first step the number of cells amounts to 10^{10} , and certainly, it is impossible to create an adequate numerical method operating with such great arrays.

One can see from (5) that the correlation dimension is determined from the ℓ dependence of the number of the set

points being within distance ℓ . One should be given by the values of $\{\ell_i\}$ and estimate for each of them so-called correlation integral (numerator of formula (5)). In Refs. [16,17] further simplifications of formula (5) are suggested. Using ergodic theorem one can make a replacement:

$$\sum_{i=1}^{M(\ell)} P_i^2 = \frac{1}{N} \sum_{j=1}^N \tilde{P}_j, \quad (6)$$

where \tilde{P}_j is the probability to find the point of the studied set not simply on the attractor but inside the hyperball of radius ℓ , with a centre at some other point of the studied set.

Further, analyzing formulae (1), (5), (6), we can show that the correlation integral $C(\ell)$ is simply equal to average number of points inside hyperballs of radius ℓ with centres at the points of the set. And for numerical calculation of the correlation dimension the following relation is used:

$$C(\ell) \sim \ell^D \quad (7)$$

Calculating the values of the correlation integral for several (≥ 3) values of ℓ , we can estimate D as a slope of the straight line connecting $C(\ell)$ and ℓ in double-logarithmic scale. Numerical calculations are carried out for a fixed series $\{\ell_j\}$ and some finite N . However there are no instructions regarding the choice of these parameters.

We'll try to overcome this drawback and, by introducing some natural scale, to remove uncertainty in the choice of $\{\ell_j\}$.

Let us replace in formula (7) ℓ by \bar{R}_K , where \bar{R}_K is the set-averaged distance to the K -th nearest neighbour (KNN); so we obtain

$$C(\bar{R}_K) \sim (\bar{R}_K)^D. \quad (8)$$

Notice that the left-hand side is equivalent to the average number of the set points being inside the hyperball with a radius equal to the average distance to the K -th neighbour, i.e. equal to the number

$$K \sim (\bar{R}_K)^D. \quad (9)$$

Hence, the modified algorithm determines D as a slope of K dependence of \bar{R}_K in double-logarithmic scale over several values of $\{K_j\}$ (we usually take $K_j = 3, 4, \dots, K$, $K \approx \sqrt{N}$; the study of N dependence of K in nonparametric estimation of probability density is presented in Refs. [18,19]).

Thus, we introduced a natural scale - average distance to the nearest neighbour - and obtained a relation between the values of N and K parameters. As will be seen below from the results of simulations, the choice of $\{K_j\}$ values contrary to $\{\ell_j\}$ is not too critical relative to the shift and spread of obtained values.

2. KNN Estimation of Probability Density. Local and Global Estimations of Dimension.

Consider nonparametric KNN estimation of probability density which is a development of simple histogram methods [20,21] :

$$P_{K,N}(X_i) = \frac{K}{N \cdot V_{K,N}(X_i)} \quad (10)$$

where $V_{K,N}(X_i)$ is a volume of \mathcal{D} -dimensional hypersphere containing the nearest to X_i representatives of the studied set:

$$V_{K,N}(X_i) = V_{\mathcal{D}} R_{K,N}^{\mathcal{D}}; \quad V_{\mathcal{D}} = \frac{\pi^{\mathcal{D}/2}}{\Gamma(\mathcal{D}/2 + 1)} \quad (11)$$

From (10) and (11) we can readily obtain (see [22]):

$$\ln R_{K,N}(X_i) = \frac{1}{\mathcal{D}} \ln K + \ln [N \cdot V_{\mathcal{D}} \hat{P}_{K,N}(X_i)]^{-1/\mathcal{D}} \quad (12)$$

Eq.(12) cannot be solved relative to \mathcal{D} , since $\hat{P}(X_i)$ depend on K . Therefore, we perform averaging of $R_{K,N}$ over the whole set according to the distribution function:

$$f_{K,X}(R) = C \mathcal{D} R^{\mathcal{D}-1} \frac{(CR^{\mathcal{D}})^{K-1}}{\Gamma(K)} \exp(-CR^{\mathcal{D}}), \quad (13)$$

where $C = NP(X)V_{\mathcal{D}}$

Replacing the mathematical expectation value by sampling average (or median) we'll obtain in the approximation of small R and large N the following equation:

$$\ln G_{K,\mathcal{D}} + \ln R_{K,N} = \frac{1}{\mathcal{D}} \ln K + C,$$

$$G_{K,\mathcal{D}} = K^{1/\mathcal{D}} \Gamma(K) / \Gamma(K + 1/\mathcal{D}), \quad (14)$$

The difference of (14) from (7) and (9) consists in the so-called iterative addition, $G_{K,d}$, which is close to zero for all K and d . Therefore we solve Eq.(14) iteratively, first assuming $G_{K,d} = 0$, and then, having obtained \hat{d}_i , we calculate G_{K,\hat{d}_i} and determine a new value of \hat{d}_{i+1} . We'll stop the iterative process, when change practically no longer takes place. Such verification of \hat{d} estimates is connected with averaging of the correlation integral.

The correlation integral equivalent to the number of the set points inside the hypersphere of radius $R_{K,N}$ is a random quantity having binomial distribution with parameter $P(X_i)$ - the probability for the point to fall into this hyperball. In the approximation of small R and large N this distribution is well-approximated by the Poisson distribution (13).

Thus, we obtained the method of estimation of dimension for the finite set of experimental events, and we'll apply it to the analysis of multiple reactions. Notice that we obtained a global estimate, i.e. the whole set is characterized by a unique number, although local differences are possible in it. From this point of view, local estimation of dimension is much more interesting to us, since in this way we'll be able to extract local inhomogeneities corresponding to various dynamical mechanisms and, possibly, to isolate resonance production on the background of invariant phase volume.

Consider Eq.(12) again. Apart from the set averaging, there is also another possibility to get linear equation for the determination of dimension. To do so, we should choose $\{K_j\}$ such that the density estimates would be very close, and hence

the dependence of $\hat{P}_{K,N}(X_i)$ on K could be ignored. Following these chosen values of $\{K_i\}$ and corresponding $\{R_{K_i}(X_i)\}$ we'll determine estimates of local dimension at a point X_i . Such estimates of density depend on dimension; it is necessary to organize the iterative procedure, i.e. for the current estimates of dimension we'll choose again the series $\{K_j\}$ corresponding to close values of densities and so on. We'll interrupt the iterative process when the value of dimension will practically no longer change. Usually 2-3 iterations turn out enough to satisfy the condition $|\mathcal{D}_{i+1} - \mathcal{D}_i| \leq 0.01$

3. Results of Monte-Carlo Simulations.

We applied our technique to many simple examples (Coch curve, Serpinsky carpet, Cantor set, etc.) and obtained estimates being in good agreement (with account of limitedness of samples and generations) with theoretical values.

The studies with Monte-Carlo simulations of multiple production events were aimed at a comparison of "pure states" - the resonance production events and the events when interactions between secondary particles are absent (I and II of Fig.1). Apart from that, the possibility of extraction of the resonance production events on the kinematic background was studied.

We generated samples according to schemes 1 and 2 with respect to resonant width and arbitrary momentum resolution. Further, by formula (14), we determined dimension of the set of points for various values of parameters K and N . Averaging was performed over 10 independent samples. As is seen from Figs. 2 and 3, the dimension criterion allows to distinguish

with high precision various dynamical mechanisms of final state production. The values of estimates are stable with respect to the choice of the method parameters and the sampling volume of the order of 200 is sufficient for reliable recognition. The errors of estimates increase with the growth of dimension, which agrees perfectly with the practice of multi-dimensional statistics [23]. The errors decrease with growing N and K , and this testifies to the method validity and to decreased influence of fluctuations with growing sample size.

Possible ways of the method utilization will be discussed in the Conclusion, while here we'll mention a relation of obtained characteristics to the number of degrees of freedom in the final state. By the known formula [24] $N = 3M - 4$, M is the number of particles in final state. For the resonance production (II) $N = 4$, non-resonance (I) $N = 8$. Of course, the possibility to recognize "pure states" is of interest, but actually at large multiplicities the final state is a mixture of various modes, and it is just necessary to extract from the background the events corresponding to nontrivial dynamical mechanisms. We may assume that in such a mixture local inhomogeneities and clusters can be observed, which have different dimension reflecting the production mechanism. Therefore, the next step in our study was determination of local dimension in a mixture of two "pure states" I and II. Fig.4 shows that the presence of resonance whose fraction decreased down to 20% clearly is an excess over background corresponding to non-resonant production according to invariant phase volume.

The iterative procedure began with the value $K=25$, then

we chose 5 median values of density (ordered statistics from 10 to 14), dimension was determined by the relevant values of $\{K_j\}$ and $\{R_{K_j}(x_i)\}$; then, with the new value of dimension we again determined densities, and so on, until the change in dimension was less than 0.01.

In this way we determined dimension for each event of the sample.

The program uses fast-sorting algorithms [25], therefore, time spent for obtaining dimension distribution is not much.

Conclusion.

We demonstrated that the proposed method of analysis of kinematic information of multiple production allows to recognize "pure" states - samples consisting of entirely background process and resonance production. Besides, the local dimension distributions allow to extract the resonance production events. Thus we can judge also on a fraction of corresponding channels of reaction. The method may be recommended for preliminary analysis of kinematic information. Further, combining it with the cluster analysis [26,27] and effective mass analysis one can determine besides the fact of existence of resonances themselves also their widths and masses. Algorithms of dimension analysis are rather simple and fast and offer an opportunity to visualize multidimensional information.

In conclusion the authors would like to express their sincere gratitude to H.R.Gulkanyan, S.G.Matinyan and G.K.Savvidy for the useful discussions; to Ts.A.Amatuni and G.Ranshall for

presenting programs. One of the authors (A.A.C.) is thankful to I.Dremin and I.Sokolov for the valuable remarks.

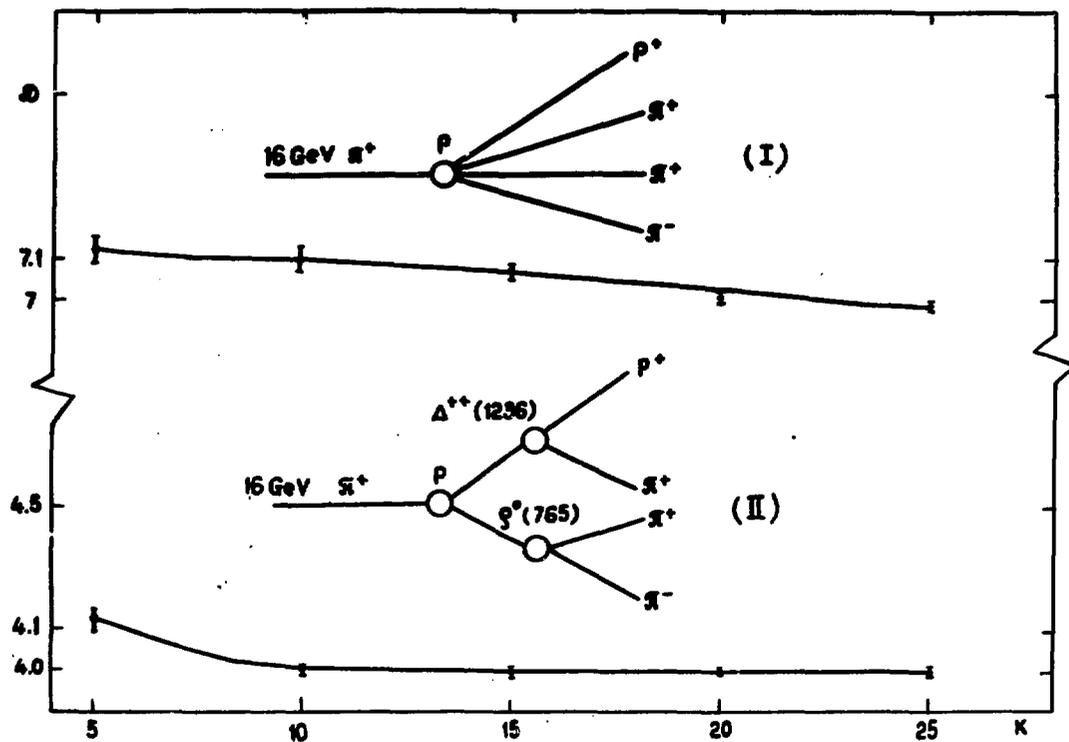


Fig.1

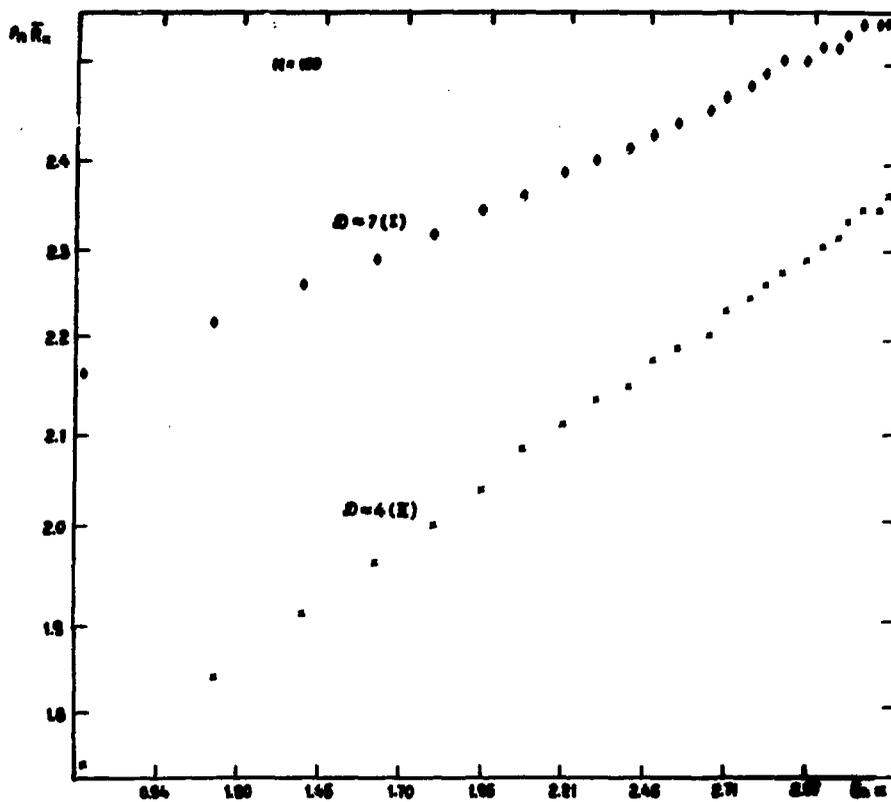


Fig.2

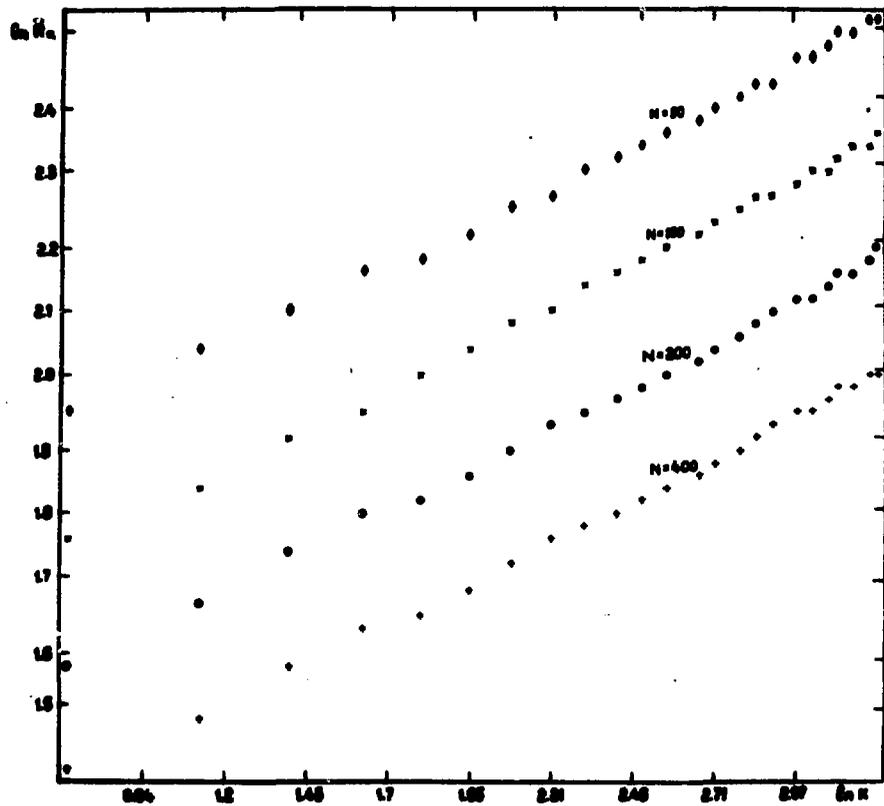


Fig.3

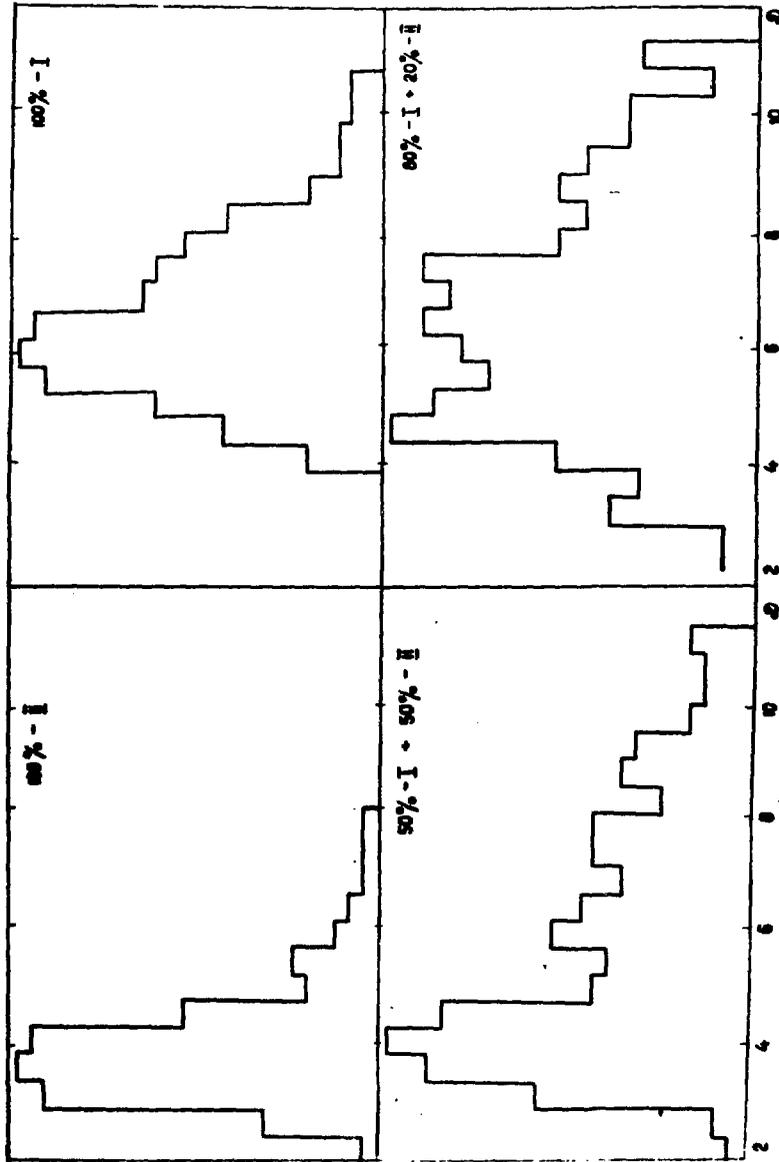


FIG. 4

Figure Captions

- Fig.1. Comparison of global estimates of correlation dimension for two versions of obtaining a given final state. Dimension of initial space is 16. The number of degrees of freedom of version I is 8, of version II is 4.
- Fig.2. Dependence of average distance to the K -th neighbour on K , by which the correlation dimension is determined.
- Fig.3. Determination of correlation dimension by different number of events of multiple production (version II).
- Fig.4. Local dimensions distribution for different proportions of the mixture of events of I and II type.

REFERENCES

1. Reni A. Probability Theory, North-Holland, Amsterdam, 1970.
2. Grassberger P., Procaccia I. Characterization of strange attractors. - Phys.Rev.Lett., 1983, v.50, p.346-349.
3. Meakin D. Scaling properties for the growth probability measure of fractal structures. - Phys.Rev.A, 1987, v.35, p.2234-2245.
4. Pagels H.R. Fractal geometry of cosmic strings and correlations among Galaxies Abel Clusters. - Phys.Rev.D, 1987, v.35, p.1141-1145.
5. Dremin I.M. Fluctuations, intermittency and fractal dimensions in multiple production. - Preprint CERN-TH.2693/87.
6. Mandelbrot B.B. The Fractal Geometry of Nature, W.H.Freeman and C^o, New York, 1982.
7. Kittel W. Summary Talk Int. Symp. on Antinucleon-Nucleon Interactions, Prague-Liblice, 1974.
8. Mandelbrot B.B. Fractals - Form, Chance and Dimensions, W.H.Freeman, San-Francisco, 1977.
9. Lichtenberg A.J., Liberman M.A. Regular and Stochastic Motion, Springer Verlag, New York, Heidelberg, 1983.
10. Eckman J.P., Ruelle D. Ergodic theory of chaos and strange attractors. - Rev.Mod.Phys., 1985, v.57, p.617-656.
11. Hausdorff F. Dimension and auberes. - Mass.Math.Ann., 1919, v.79, p.157-179.
12. Kolmogorov A.N., Tikhomirov B.V. E-entropies and capacities of sets in functional species. - Uspekhi Mat. Nauk, 1959, v.14, issue 2(86), p.3-86.

13. Hentschel H.G., Procaccia I. The infinite number of generalized dimensions of fractals and strange attractors. - Physica, 1983, v.8D, p.435-444.
14. Young L.S. Dimension entropy and Lyapunov exponents in differentiable dynamic systems. - Physica, 1984, v.124A, p.639-646.
15. Mandelbrot B.B. Fractals and turbulence: Attractors and dispersions. - Lect. Not. in Math., 1977, N.615, p.83-93.
16. Pawelzik K., Shuster H.S. Generalized dimensions and entropies from a measured time series. - Phys.Rev.A, 1987, v.35, p.481-484.
17. Caputo J.G., Atten P. Metric entropy: An experimental means for characterizing and quantifying chaos. - Phys.Rev.A, 1987, v.35, p.1311-1315.
18. Galfayan S.H., Chilingaryan A.A. Calculation of the Bayes Risk by estimation of probability density function by KNN method. - Stat. Prob. of Control, Vilnius, 1985, v.66, p.66-78.
19. Chilingaryan A.A. Statistical decisions under nonparametric a priori information. - Preprint YERPHI, 819(49), Yerevan, 1985.
20. Parzen E. On estimation of a probability density function and mode. - Ann. Math. Stat., 1962, v.33, p.1065-1076.
21. Tapia R.A., Thompson T.R. Nonparametric Probability Density Estimation, The John Hopkins University Press, Baltimore and London, 1978.
22. Pettis K.W., Baily T.A., Tain A.K., Dubes R.C. An intrinsic dimensionality estimator from near-neighbour information.-

- IEEE Trans. on Pattern Analysis, 1979, PAMI-1, p.25-38.
23. Meisel W.S. Computer Oriented Approaches to Pattern Recognition, Academic Press, New York and London, 1972.
 24. Byckling E., Kajantie K. Particle Kinematics, John Wiley and Sons, London-New York, 1973.
 25. Braams B. CERN Pool, Library Entry-G1003.
 26. Gelsema E.S. Description of an interactive clustering techniques and its application. - Preprint CERN DD/74/16, 1974.
 27. Schiller H. Investigation of the multiparticle final states by the cluster analysis method in the middle energy reactions. - Particles and Nuclei, 1980, v.11, N.1, p.182-235.

The manuscript was received 18 February 1988

С.Х. АРУТЮНЯН, А.А. ЧИЛИНГАРЯН

О ВОЗМОЖНОСТИ АНАЛИЗА МНОГОМЕРНОЙ КИНЕМАТИЧЕСКОЙ ИНФОРМАЦИИ
С ПОМОЩЬЮ КЭС ОЦЕНОК РАЗМЕРНОСТИ

(на английском языке, перевод Э.Н. Асланян)

Редактор Л.П. Мукаян

Технический редактор А.С. Абрамян

Подписано в печать 7/VI-88г.
Офсетная печать. Уч. изд. л. 1,0
Заказ тип. № 273

ВФ-03147

Формат 60x84/16
Тираж 299 экз. Ц. 15к
Индекс 3624

Отпечатано в Ереванском физическом институте
Ереван 36, Маркаряна 2

The address for requests:
Information Department
Yerevan Physics Institute
Markaryan St., 2
Yerevan, 375036
Armenia, USSR

индекс 3624



ЕРЕВАНСКИЙ ФИЗИЧЕСКИЙ ИНСТИТУТ