Nuclear Instruments and Methods in Physics Research A281 (1989) 388–392
North-Holland, Amsterdam

ON THE POSSIBILITY OF A MULTIDIMENSIONAL KINEMATIC INFORMATION ANALYSIS BY MEANS OF NEAREST-NEIGHBOUR ESTIMATIONS OF DIMENSIONALITY

A.A. CHILINGARIAN and S.Kh. HARUTUNYAN

Yerevan Physics Institute, Alikhanian Brothers St. 2, Yerevan 375036, Armenia, USSR

Received 14 September 1988 and in revised form 9 January 1989

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

1. Introduction

Recently, great success in the description of complex system behaviour was achieved by using geometrical representations. The generalized dimensionality originally introduced by Renyi [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] up to the star cluster [4] and quark–gluon plasma [5].

On the other hand, the development of Mandelbrot's ideas about the fractal character of Nature [6] also brought a new understanding of physical experimental data.

The kinematic information on the high-multiplicity reactions sharply increases and, simultaneously, the detection of so far unknown mechanisms of production of a given final state becomes more difficult [7]. The effective mass distribution does not allow any definite conclusions.

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

The essential inhomogeneity and complexity of the event patterns in phase space gave us the idea of using a fractal approach for the analysis of multiple production. In a wide sense a fractal set is a set whose structure is related to dimensionality [8]. Fractal analysis proves to be useful every time when the system's behaviour is characterized by attractivity. That is, final states are grouped in some bounded 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 ways to calculate them for finite sets [10], most of which go back to the first generalization of the dimensionality notion by Hausdorf [11]. However, for the cases important to physical experiments, most of these definitions are equivalent; therefore we shall prefer the methods allowing one to consider large dimensions of the initial space.

Strict mathematical definitions of dimensionality as well as references to the basic works can be found in ref. [12].

Highly useful proved to be the approach worked out by Procaccia et al. and Young [13,14], which allows the generalization of some of the most popular definitions of dimensionality and creates a numerical method of calculation.

Note that the aim of the fractal approach is not to provide us with a ready theory but to formulate empirical facts in terms of geometry [15] for a subsequent comprehensive analysis.

2. Correlation dimension calculation

Procaccia showed that there exist an infinite number of generalized dimensions characterizing an attractor:

$$D_q = \frac{1}{q-1} \lim_{l \to 0} \ln \sum_{i=1}^{M(l)} P_i^q / \ln l, \qquad (1)$$

in the d-dimensional initial space where an embedded attractor is divided into $M(l)$ cubes (boxes, cells,

0168-9002/89/$03.50

bins, $\cdots$), and in each of them a probabilistic measure P_i is determined; the cube volume equals l^d, where q is an arbitrary real number. One can easily show that for $q \to 0$ the generalized dimension coincides with the self-similarity dimension D_s:

$$D_s = \frac{\ln(M_{k+1}/M_k)}{\ln(l_{k+1}/l_k)}, \tag{2}$$

where M_k is the number of self-similar objects produced at the Kth scale fragmentation step. The self-similarity dimension in its turn is closely related to the Hausdorf dimension D_h:

$$D_h \equiv D_{q\to 0} = -\lim_{l\to 0}\lim_{N\to\infty} \ln M(l)/\ln(l), \tag{3}$$

where N is the number of points on the attractor, i.e. the sample (set) size.

Practically, the dimension is determined as a slope of a straight line connecting some l values with $M(l)$ values on a double-logarithmic scale. To do so, one should of course, start with a series $\{l_i\}$, $i = 1, 2, \cdots k$, $k \geq 3$ and calculate the relevant series of $\{M(l_i)\}$ – the number of cells of the volume l_i^d including all the points of the studied set.

At $q \to 1$ the generalized dimension reduces to an information one:

$$\sigma \equiv D_{q\to 1} = \lim_{l\to 0}\lim_{N\to\infty} \sum_{i=1}^{M(l)} P_i \ln P_i/\ln l. \tag{4}$$

Most important for the applied cases is the correlation dimension D_c corresponding to the case $q = 2$:

$$D_c = D_{q=2} = \lim_{l\to 0}\lim_{N\to\infty} \sum_{i=1}^{M(l)} P_i^2/\ln l. \tag{5}$$

The correlation dimension is important, firstly, because it characterizes the local structure of the attractor, and secondly, because, as will be seen further on, it can readily be calculated for dimensions of initial space $d \gg 2$. On the other hand, the algorithm of direct counting of cells is rather time-consuming and is applicable only for the cases when $d \leq 2$. 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 it is impossible to develop an adequate numerical method operating with such a bulk of information.

One can see from eq. (5) that the correlation dimension is determined from the l-dependence of the number of set points being within distance l. One should start out with the values of $\{l_i\}$ and calculate for each of them the so-called correlation integral $C(l)$ – the numerator of eq. (5). In refs. [16,17] some simplifications of the method for correlation dimension calculation are suggested. Using the ergodic theorem one can make a replacement:

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

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

Further, analyzing eqs. (1), (5) and (6), one can show that the correlation integral $C(l)$ is simply equal to the mean number of points inside the hypersphere of radius l centered at the set point. And finally the correlation dimension can be calculated from the l-dependence of the correlation integral:

$$C(l) \sim l^{D_c}. \tag{7}$$

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

We shall try to overcome this drawback by introducing some natural scale. Let us replace l by $\bar{R}_K$ in eq. (7) – the sample-averaged distance to the Kth nearest neighbour (KNN):

$$C(\bar{R}_K) \sim \bar{R}_K^{D_c}. \tag{8}$$

Notice that the left-hand side is equivalent to the mean number of sample points inside a hypersphere with radius equal to the average distance to the Kth neighbour, i.e. it equals the number K, so:

$$K \sim \bar{R}_K^{D_c}. \tag{9}$$

Hence, the modified algorithm defines D_c as the slope of the K-dependence of $\bar{R}_K$ on a double-logarithmic scale. (We usually take $\{K_j\} = 1, 2, \cdots \mathcal{K}$, $\mathcal{K} = \sqrt{N}$; the study of the N-dependence of $\mathcal{K}$ in the estimation of the probability density is presented in refs. [18,19].)

Thus, we introduce a natural scale – the average distance to the nearest neighbours. As will be seen below from the simulation, the choice of $\{K_j\}$ values, in contrast to $\{l_j\}$, is not too critical for the dimension estimation.

3. KNN estimation of probability density; local and global dimensionality

Consider the KNN estimation of the probability density which is a development of the well-known histo-

gram method [21,22]:

$$p_k(x_i) = \frac{k}{NV_k(x_i)}, \tag{10}$$

where $V_k(x_i)$ is the volume of a d-dimensional sphere containing the K nearest to x_i representatives of the set studied (x_i belongs to the same set),

$$V_k(x_i) = V_d R_k^d; \quad V_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}, \tag{11}$$

where R_k is the distance to the Kth nearest neighbour of x_i, $\Gamma(\,)$ is the gamma function. From eqs. (10) and (11) we can readily obtain (see ref. [22]):

$$\ln R_k(x_i) = \frac{1}{d}\ln K + \ln\left[NV_d p_k(x_i)\right]^{-1/d}. \tag{12}$$

Eq. (12) cannot be solved relative to d, since the estimate of $p(x_i)$, as one can see from eq. (10), depends on K. Therefore, one can average R_k over the whole sample, according to the distribution function:

$$f_{k,N}(R) = CdR^{d-1}\frac{(CR^d)^{k-1}}{\Gamma(k)}\exp(-CR^d), \tag{13}$$

where $C = Np(x)V_d$.

In the approximation of small R and large N we obtain the following equations:

$$\ln G_{k,d} + \ln \bar{R}_k = \frac{1}{d}\ln K + \text{const}, \tag{14}$$

$$G_{k,d} = K^{1/d}\Gamma(k)/\Gamma(k+1/d),$$

where $\bar{R}_k$ is the sample-averaged distance to the Kth nearest neighbour and "const" is independent of K.

The difference of this equation from the previous ones (eqs. (7) and (9)), obtained by a completely different approach, consists in the so-called iterative addition $G_{k,d}$, which is close to zero for all K and d. Therefore, we solve this equation iteratively, first assuming $G_{k,d} = 0$, and then, having obtained d_i, we calculate $G_{k,d}$ and determine the value of d_{i+1}. We shall stop the iterations when d is practically constant.

Such a verification of d estimates is connected with averaging of the correlation integral. The correlation integral – the number of sample points inside the hypersphere of fixed radius – is a random variable belonging to the binomial distribution with parameter $P(x)$ (the probability for the sample point to fall within this hypersphere).

Thus, we obtain the method of dimension estimation for a finite set of experimental events, and we shall apply it to the analysis of multiple production.

It should be noted that our estimate is a global estimate, i.e. the whole sample is characterized by one number, though local differences are possible. From this point of view, local dimensionality is much more interesting, since we shall be able to detect local inhomogeneities corresponding to various dynamical mechanisms and, possibly, to isolate resonance production.

Consider eq. (12) again. Apart from sample averaging, there is also another way to get a linear equation for determining the dimension. For this, one must choose the $\{K_j\}$ series such that the density estimates are very close and, hence, the dependence of $p_k(x)$ on K can be ignored. Following these chosen values of $\{K_j\}$ and the corresponding $\{R_{k_j}(x_i)\}$, one can carry out the estimate of the local dimension at a point x_i.

As the density estimates depend on dimension, it is necessary to settle an iterative procedure, i.e. for the current dimension value choose again the series $\{K_j\}$, corresponding to the values of density close to each other, and so on, and interrupt the iterations when the value of the dimension is practically constant. Usually, 2–3 iterations are enough to satisfy

$$|d_{i+1} - d_i| \le 0.01.$$

4. Results of the simulations

We applied the developed technique to determine the dimension of many standard sets (Coch curve, Serpinsky carpet, Cantor set, etc.), and obtained estimates in good agreement (taking into account the limitedness of generated samples) with theoretical values.

The simulations of multiple production were aimed at a comparison of "pure states" – resonsance production events and events when interactions between secondary particles are absent. Apart from that, the possibility of indication of resonance production events was studied. Two channels of 16 GeV π^+ hadron production were considered:

$$\begin{aligned} &(1)\quad \pi^+ p \to p\pi^+\pi^+\pi^- \\ &(2)\quad \pi^+ p \to \rho^0(765)\,\Delta^{++}(1236), \quad \rho^0 \to \pi^+\pi^-, \quad \Delta^{++} \to p\pi^+ \end{aligned} \tag{15}$$

We generate samples according to reactions (1) and (2) with account of the resonant width and a given momentum resolution. Further, by eq. (14), we determined the dimension for various values of $\mathcal{K}$ and N. Averaging was performed over 10 independent samples of fixed size. As is seen from figs. 1, 2 and 3, the dimension criterion allows us to distinguish with high precision between various dynamical mechanisms of final state production. The values of the estimates are stable with respect to the choice of method parameters, and the sample size of 200 seems to be sufficient for a reliable recognition. The errors of the estimates increase with the dimension, which agrees with the practice of multivariate analysis [23]. The errors decrease with N and $\mathcal{K}$, and this testifies the consistency of the method

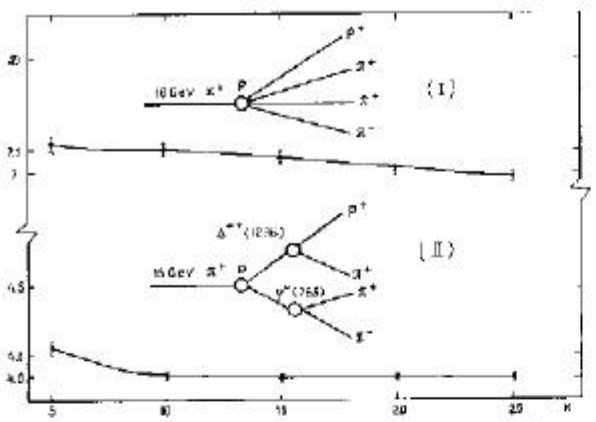

Fig. 1. Comparison of global estimates of correlation dimension for two ways of obtaining the given final state. The initial space dimension is 16. The number of degrees of freedom of the first channel is 8 and that of second 4.

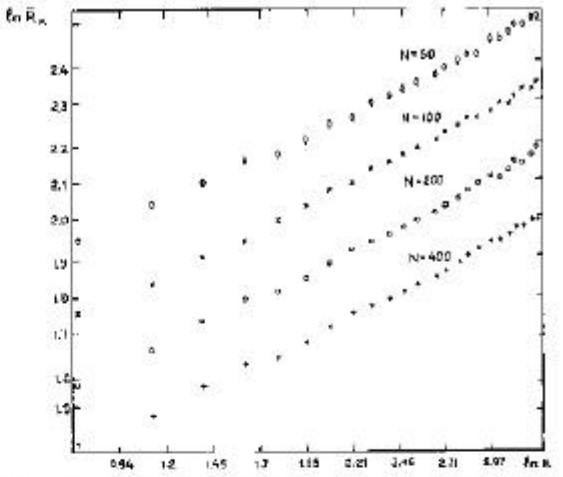

Fig. 3. Determination of dimension for different numbers of events of multiple production.

and the decreased influence of fluctuations as the sample size grows.

Possible ways of utilisation of the method will be discussed in the conclusion, while here we shall mention a relation of the obtained characteristic to the number of degrees of freedom $\mathcal{N}$ in the final state. By the well-known formula [24] $\mathcal{N}=3m-4$ (m is the number of particles in the final state), for the nonresonance production (1) $\mathcal{N}=8$. For reaction (2), where there are two additional conservation equations for each secondary vertex, $\mathcal{N}=4$.

Of course, the possibility to recognize the "pure states" is of interest, but at large multiplicities the final state is a mixture of various modes, and it is necessary to extract from the background the events corresponding to nontrivial dynamical mechanisms. We may assume that in such a mixture the local inhomogeneities and clusters of different dimensions can be observed that characterize the production mechanism. Therefore, the next step in our studies was the determination of local dimension in a mixture of two "pure states". Fig. 4 shows that the presence of a resonance whose fraction decreased down to 20% is clearly seen as an excess in the local dimension histogram (the dimensions were calculated at each point of the studied sample).

The iterative procedure for local dimension determination began with the value of $\mathcal{N}=25$, then we chose 5 median values of density (ordered statistics from 10 to 14), and determined the dimension by the relevant values of $\{K_{1/l}\}$ and $\{R_{K_{1/l}}\}$, where brackets indicate that the K values are taken corresponding to the ordered sequence of the density estimates. Then, with the calculated value of the dimension we again determined the density sequence, corresponding to dif-

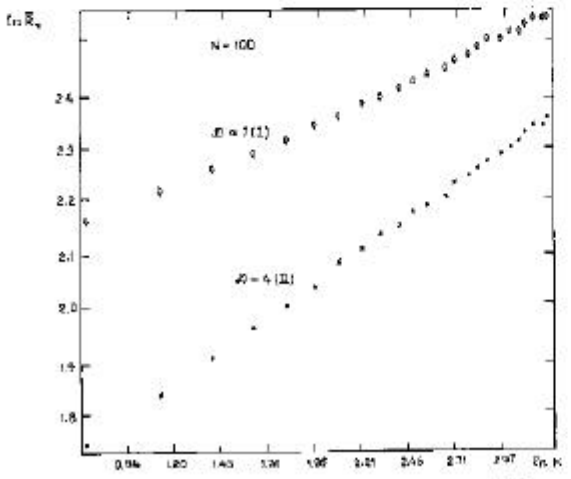

Fig. 2. Dependence of average distance to the K th neighbour, $\bar{R}_K$, on K, by which the dimension is calculated.

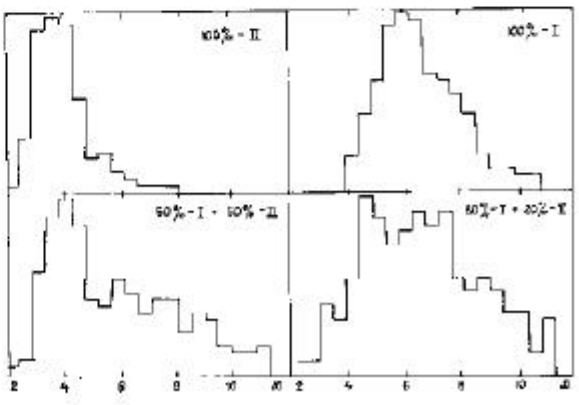

Fig. 4. Local dimension distribution for different proportions of type (1) and (2) events.

ferent K values, chose the median values and relevant $\{K_{|J|}\}$ values, and so on, until the change in the dimension estimates was less than 0.01.

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

The program uses fast-sorting algorithms [25], therefore the time spent in obtaining the dimension distribution is not too long.

5. Conclusion

We demonstrated that the proposed method of analysing kinematic information allows one to recognize "pure states" samples consisting of a complete background or resonance production. Besides, the local dimension distribution allows one to extract the "resonance" events. Thus, we can also judge about the branching ratio of the reaction studied. The method can be recommended for preliminary analysis of kinematic information. Further, combining it with cluster analysis [26,27] and effective mass analysis, one can determine also the existence of the resonances themselves, and also their widths and masses. The algorithms for dimension analysis are rather simple and fast and offer an opportunity to visualize multidimensional information.

Acknowledgements

The authors would like to express their sincere gratitude to H.R. Gulkanyan, S.G. Matinyan and G.K. Savvidy for useful discussions; to Ts.A. Amatuni and H. Ranshali for presenting programs. One of the authors (A.A.C.) is thankful to I.M. Dremin and I.M. Sokolov for valuable remarks.

References

[1] A. Renyi, Probability Theory (North-Holland, Amsterdam, 1970).
[2] P. Grassberger and I. Procaccia, Phys. Rev. Lett. 50 (1983) 346.
[3] D. Meakin, Phys. Rev. A35 (1987) 2234.
[4] H.R. Pagels, Phys. Rev. D35 (1987) 1141.
[5] I.M. Dremin, Preprint CERN-TH 2693 (1987).
[6] B.B. Mandelbrot, The Fractal Geometry of Nature (W.H. Freeman, New York, 1982).
[7] W. Kittel, Summary Talk, Int. Symp. on Antinucleon–Nucleon Interactions, Prague–Liblice (1974).
[8] B.B. Mandelbrot, Fractals – Form, Chance and Dimension (W.H. Freeman, San Francisco, 1977).
[9] A.J. Lichtenberg and M.A. Liberman, Regular and Stochastic Motion (Springer, New York, Heidelberg, 1983).
[10] J.R. Lekman and D. Rulle, Rev. Mod. Phys. 57 (1985) 617.
[11] F. Hausdorf, Mass. Math. Ann. 79 (1919) 157.
[12] A.N. Kolmogorov and B.V. Tikhomirov, Uspekhi Math. Nauk 14 (1959) 3.
[13] H.G. Hentschel and I. Procaccia, Physica 8D (1983) 435.
[14] L.S. Young, Physica 12A (1984) 639.
[15] B.B. Mandelbrot, Lect. Notes Math. 615 (1977) 83.
[16] K. Pawelzik and H.S. Shuster, Phys. Rev. A35 (1987) 481.
[17] J.G. Caputo and P. Atten, Phys. Rev. A35 (1987) 1311.
[18] A.A. Chilingarian and S.Ch. Galfayan, Stat. Prob. Control, Vilnius 66 (1985) 66.
[19] A.A. Chilingarian, to be published in Comput. Phys. Commun (1989).
[20] E. Parzen, Ann. Math. Stat. 33 (1962) 1065.
[21] R.A. Tapia and T.R. Thompson, Nonparametric Probability Density Estimation (The John Hopkins University Press, Baltimore and London, 1978).
[22] K.W. Pettis, T.A. Baily, A.K. Tain and R.C. Dubes, IEEE Trans. Pattern. Anal. Machine Intelligence PAMI-1 (1979) 25.
[23] W.S. Meisel, Computer Oriented Approaches to Pattern Recognition (Academic Press, New York and London, 1972).
[24] E. Byckling and K. Kajantee, Particle Kinematics (Wiley, London, New York, 1973).
[25] B. Braams, CERN POOL library, Entry - G1003.
[26] E.S. Gelsema, Preprint CERN DD 74 (1974).
[27] H. Schiller, Part. Nucl., Dubna 11 (1980) 182.