Big Bang–Like Phenomenon in Multidimensional Data

Marcel Jiřina
Institute of Computer Science As CR, v.v.i.
Pod Vodárenskou Věží 2, 18207 Prague 8, Czech Republic
marcel@cs.cas.cz

ABSTRACT

Notion of the Big Bang in Data was introduced, when it was observed that the quantity of data grows very fast and the speed of this growth rises with time. This is parallel to the Big Bang of the Universe which expands and the speed of the expansion is the larger the farther the object is, and the expansion is isotropic. We observed another expansion in data embedded in metric space. We found that when distances in data space are polynomially expanded with a proper exponent, the space around any data point displays similar growth that is the larger the larger is the distance. We describe this phenomenon here on the basis of decomposition of the correlation integral. We show that the linear rule holds for logarithm of distance from any data point to another and proportionality constant is the scaling exponent, especially the correlation dimension. After this transformation of distances the data space appears as locally uniform and isotropic.

KEYWORDS

Big Bang, scaling, correlation dimension, expansion of distances, polynomial transformation.

1 INTRODUCTION

An extreme growth of data processed was named, for different reasons and in many institutions, the Big Bang in Data. It characterizes this extreme growth that has several aspects. There are automatically generated data, arising from uninterrupted monitoring of processes and objects. Compared to history, there is rigorous data collection. Also, there are many repeated data especially redundancies and duplicates. Also, dimensionality of data grows. The number of data items needed for description of a phenomenon or an object, and some "items" grow so that e.g. text description is substituted by a picture. There is an analogy with the Big Bang of the Universe.

The Big Bang of the Universe describes that our world has arisen from initial singularity at the moment of the Big Bang 14 billion years ago. Since then the Universe has been expanding linearly. Observer in the Universe can see that objects in the Universe, galaxies, recede so that the farther galaxy is, the faster is the receding. Astronomers know this thanks to the Doppler Effect. In 1929 Edwin Hubble found a constant with which the Universe expands. He explained observed red shift in spectra, by the Doppler effect and stated a constant of expansion, now called the Hubble constant, that is $H_0 = 74.8$ (meters per hour)/(light year). The most important fact is that the expansion is linearly proportional to distance,

$$v_G = H_0 d_G \tag{1}$$

where v_G is velocity of the receding, and d_G is the distance of an object G (galaxy). Of great importance is the fact that this expansion is observable from any place in the Universe. An observer in any place of the Universe can see all objects receding in the same way.

In this work we first present some known facts or effects that appear when studying multi-dimensional space. Then we use a nonlinear transformation that changes relative distances of data points so that it appears that for logarithm of distances from any point of the data set a linear rule holds that is very similar to Hubble's law of expansion of the Universe. Its important feature is

that it is also isotropic. We show here that using this transformation of distances data space around any point appears locally uniform and also that concentration phenomenon can be suppressed.

2 MULTIVARIATE AND MULTIDIMENSIONAL DATA

Before we speak about multidimensional data we must differentiate between multivariate and multidimensional.

We suppose that data is real numbers and any object is described with n reals. An object is called in many different ways as sample, observation, event, experiment or the like. Each of n reals has specific meaning and unit in which it is given (measured, recorded) and it is called a feature, eventually an attribute, item. Moreover, each feature can have its minimum and maximum, i.e. it has its support. Cartesian product of supports forms a data space. Obviously the data space is immerged into R^n. In R^n as well as in the data space n ordered feature values can be understood as a vector that originates at point $(0, 0, \dots 0)$ and ends at point P whose position is given by that n numbers discussed now.

At this moment we cannot state distance of point x from origin $(0, 0, \dots 0)$ or mutual distance between two points x_1 and x_2. What is missing is a metrics. The straightforward way is to use the Mahalanobis metrics. But we would like to introduce a class of L_p metrics, esp. the Euclidean and Manhattan metrics. What causes problems is the fact that each feature is measured in different units. Some simple transformation is needed that makes data "dimensionless" or with the same measure, usually with zero mean and unit standard deviation. Using this normalization the Euclidean metrics appears to be a special case of Mahalanobis distance with covariance matrix substituted with unit matrix. By defining metrics and using normalization above when needed, we change R^n into the metric space and transformed data space into data space with metrics, and we can speak about multidimensional data.

3 PROPERTIES OF MULTIDIMENSIONAL DATA

In multidimensional data space one can find several observations. All observations rely on the Euclidean metrics, eventually on the L_p metrics with small p.

3.1 Concentration Phenomenon

It is informally stated in statistics that "A random variable that depends in a Lipschitz way on many independent variables (but not too much on any of them) is essentially constant" [1].

In geometry of the multidimensional space let us have a ball with uniformly spread points inside. Then it appears that most of the points lie near the surface of the ball forming some peel, with nearly no points inside. The higher is the dimension, the thinner the peel. It is easy to compute thickness of the peel in which 90 per cent of the points lie. Simply, under the assumption of uniformity, what is thickness of peel that contains 90 per cent of the ball's volume?

Table 1. The "peel" thickness in the unit ball containing 90 % of volume.

dim	peel thickness
1	0.9
2	0.6838
3	0.5358
4	0.4377
5	0.3690
10	0.2057
20	0.1087
50	0.0450
100	0.0228
200	0.0114
500	0.0046
1000	0.0023
2000	0.0012
5000	0.00046
10000	0.00023
100000	2.3E-05

ISBN: 978-0-9891305-5-4 ©2014 SDIWC

Table 1 shows that in higher dimensions the peel is rather thin. Thus, it seems that nearly all the points lie on the surface of the ball and generally a "cloud" of data appears to be nearly empty and all data points are on its surface. This leads to other phenomena that have the same origin. Note that when the L_{inf} metrics is used, this effect does not appear.

3.2 Random triangles in a space

F. Murtagh [2] found that randomly generated triangles in high dimensional spaces tend to be equilateral and isosceles. More exactly, let there be a tolerance *d*. A triangle is called equilateral, if absolute value of two edges length difference is smaller than d. A triangle is called isosceles, if absolute value of the largest difference of length of any two edges is smaller than d. The "strong triangular inequality" or ultrametric inequality $d(x, z) \leq \max\{d(x, y), d(y, z)\}$ for any triplet x, y, z is also studied. Results according to [2] are given in Table 2. It is seen that for dimension n going to infinity, the percentage of equilateral triangles among all randomly generated triangles approaches 100 %, and also percentage of cases when ultrametric inequality holds approaches 100 %.

3.3 Ball/Cube Volume Relation

Here we discuss the problem how large is the volume of unit diameter ball in a unit cube in n dimensional Euclidean space in dependence on space dimensionality.

The volume of a unit cube is one volume unit, say m^n, e.g. one meter squared in two dimensional space, a plane, one cubic meter in three dimensional space and so on. It appears that a unit ball in a two-dimensional space is a circle with area 0.785398 meters squared. Data for other dimensions is shown in Table 3 and in Figure 1.

It is seen that ball in a cube is rather tiny in higher dimensions. It occupies less than 2.5 % of the cube's volume in ten-dimensional space, and 10^{-10} in 24-dimensional space.

In n-dimensional space with the Manhattan metrics the ball/cube volume ratio is smaller than in Euclidean space. On the other hand, this ratio is equal to one for any dimension when the L_{inf} metric is used.

Table 2. (According to [2].) Results, based on 300 sampled triangles from triplets of points. For uniform, the data are generated on $[0, 1]^m$; hypercube vertices are in $\{0, 1\}^m$, and for Gaussian on each dimension, the data are of mean 0, and variance 1. Dimen. is the embedding dimensionality. Isosc. is the number of isosceles triangles with small base, as a proportion of all triangles sampled. Equil. is the number of equilateral triangles as a proportion of triangles sampled. UM is the proportion of ultrametricity-respecting triangles (= 1 for all ultrametric).

No. points	Dimen.	Isosc.	Equil.	UM
Uniform				
100	20	0.10	0.03	0.13
100	200	0.16	0.20	0.36
100	2000	0.01	0.83	0.84
100	20000	0	0.94	0.94
Hypercube				
100	20	0.14	0.02	0.16
100	200	0.16	0.21	0.36
100	2000	0.01	0.86	0.87
100	20000	0	0.96	0.96
Gaussian				
100	20	0.12	0.01	0.13
100	200	0.23	0.14	0.36
100	2000	0.04	0.77	0.80
100	20000	0	0.98	0.98

Table 3. Ratio of volume of a ball in the cube in dependence on dimension.

n	Ball/Cube
1	1
2	0.785398
3	0.523599
4	0.308425
5	0.164493
6	0.080746
7	0.036912
8	0.015854
9	0.006442
10	0.00249

ISBN: 978-0-9891305-5-4 ©2014 SDIWC

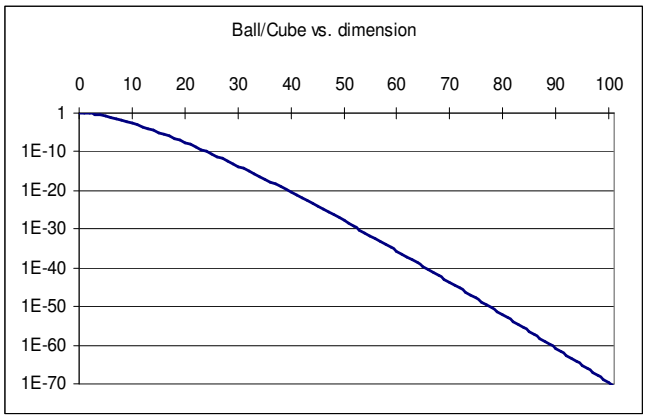

Figure 1. Ratio of volume of a ball in the cube in dependence on dimension for dimension up to 100.

3.4 Edge Effect

The edge effect is defined so that even position of the first nearest neighbor is influenced by the presence of a margin cluster of data points and outer "empty space". Data form some cluster in a metric space. Suppose the Euclidean metrics and data points uniformly spread inside. For a point inside the cluster also the largest ball of some diameter lies whole inside the cluster. Then, if the nearest neighbor lies inside this ball its position is not influenced by the boundary of the cluster. Conversely, when it lies outside such a ball, it can lie in a part of space only, and its distance is on average larger than in the previous case. A position of the second and other neighbors is influenced this way as well. Let us have a model with a cluster in the form of a hypercube of some dimension with data points randomly and uniformly spread inside. In Figure 2 it is shown that already for 100 points the position of the first nearest neighbor is influenced when space dimension is 8. In dimension 17 even when there are one million data points in the cube, the position of the first neighbor is influenced. In Figure 2 index k means the order number of neighbor. The higher dimension and the lesser the number of data points, the worse.

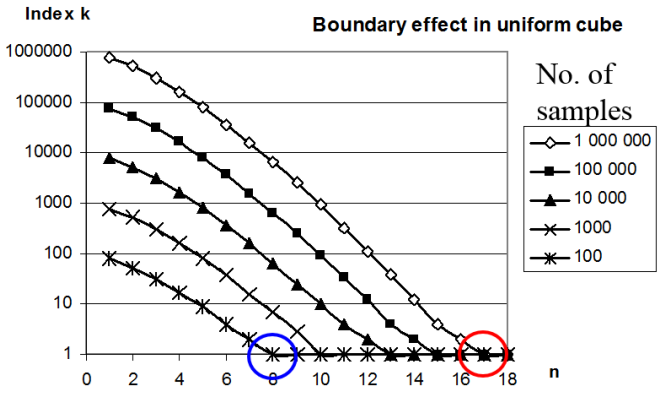

Figure 2. The order number of a neighbor whose position is affected by boundary effect.

An opponent can say that data never form a cube in R^n, but a much rounded cloud. That is true, but on the other hand, data in practice never form a ball and it is easy to show that the observation above is pessimistic but not too far from reality.
All phenomena mentioned are the same for any observer or point x. The true position, coordinates, of point x do not influence these findings.

4 BIG BANG - LIKE PHENOMENON IN DATA

The last observation is that all phenomena described are the same for any observer or point x. This is reminiscent of the fundamental observation in the Big Bang of the Universe: The expansion of the Universe is the same for any observer, i.e. it does not depend on his position or direction of observation. The space is isotropic.

4.1 Correlation Dimension

There is another observation mentioned already by Mandelbrot [3], that any practical data possess some form of scaling, i.e. they may look like arising from the fractal process. Of course, scaling does not imply a fractal, but fractal implies scaling.
As for data in a metric (Euclidean) space, one can consider individual data as points. We can compute distances between all pairs of points and construct an empirical distribution function of these distances. Thus, one can introduce a correlation integral according to [4] as follows.

$$C_I(r) = \lim_{N\to\infty} \frac{1}{N^2} \times \{number of\ pairs(i,j) : \|X_i - X_j\| < r\}$$

.

In a more comprehensive form one can write

$$C_I(r) = \Pr(\|X_i - X_j\| < r)\,.$$

The correlation integral can be rewritten in the form

$$C_I(r) = \lim_{N\to\infty} \frac{1}{N(N-1)} \sum_{1\le i<j\le N} h(r - \|X_j - X_i\|) \quad , \qquad (2)$$

where $h(.)$ is Heaviside step function. From it

$$\nu = \lim_{r\to\infty} \frac{\ln C_I(r)}{\ln r} \quad .$$

Grassberger and Procaccia [4] have shown that the correlation integral in log-log coordinates has a linear course, as shown in Figure 3.

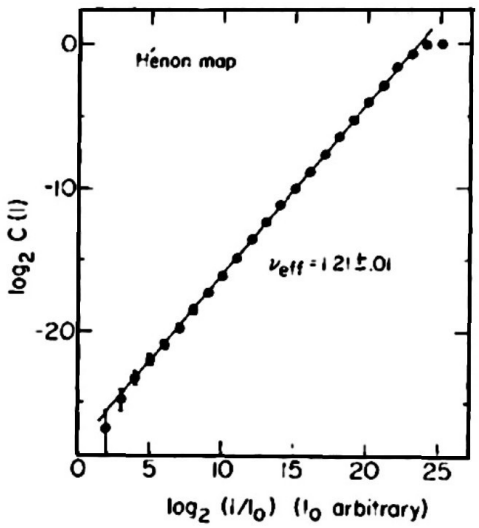

Figure 3. Correlation integral of the Hénon map according to Grassberger and Procaccia [4].

This is a typical behavior: an essential part of the correlation integral in log-log coordinates has a linear course. In the linear part then it holds

$$\ln P(l<d) = C + \nu \ln d, \qquad (3)$$

where C is a constant, and ν is the so-called correlation dimension, and d is distance between two points. More exactly, it holds

$$\nu = \lim_{d\to 0+} \frac{\ln P(l<d)}{\ln d} \quad .$$

Relation (3) holds exactly for some $d < d_0$ and then it is often used for estimating the correlation dimension from data [8], [9], [10].

4.2 Distribution Mapping Function

Here we first single out one important notion and then we will show the decomposition.

We call the dependence of the percentage of neighbor points on the distance the distribution mapping function $D(x, r)$. The distance of neighbors r is a random variable and the distribution mapping function is, in fact, the distribution function of neighbor's distances. We can use some transformation of distance r to another variable z and we also call the dependence of the percentage of neighbor points on this variable $D(x, z)$ a distribution mapping function as it gives, in fact, the same information. We call a derivative of the $D(x, r)$ or $D(x, z)$ according to r or z respectively, the distribution density mapping function $d(x, r)$ or $d(x, z)$ [11].

Intuitively, Fig. 4 gives illustration of mapping functions; "Pure" means theoretical dependence, "True" means real data. From top to bottom:

In Fig. 4a the dependence of order of near point on its distance from the query point is shown. Note that the (first) nearest neighbor is rather far from the query point and that distances of farther points from the query point differ slightly. Positions of the several furthest points are influenced by the boundary effect.

In Fig. 4b the same dependence is shown in logarithmic scale; linear dependence appears here. Note the slope that is $q = 2.5$.

In Fig. 4c the dependence of percentage of points (instead of count) on distance to the q-th power is shown. In fact, this graph is the distribution function of variable $z = r^q$. Again, a linear dependence appears. It means a uniform distribution (with the exception of several farthest points). It also appears that radii r_i , $i = 1, 2, \ldots$ grow proportionally to the q-th root of index i $r_i \approx \sqrt[q]{i}$.

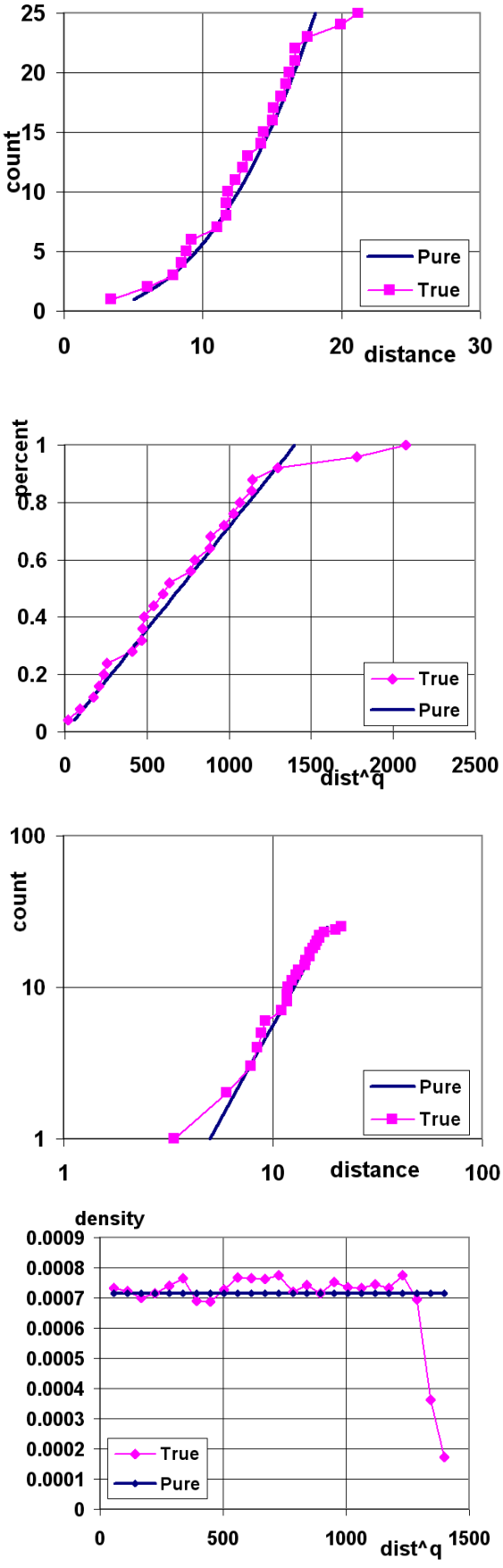

Fig. 4. Illustration of mapping functions; "Pure" means theoretical dependence, "True" means real data.
From top to bottom:

a. Dependence of order number of near point on its distance from the query point.
b. The same dependence in logarithmic scale.
c. Dependence of percentage of points (instead of count) on distance to the q-th power.
d. Density (histogram) of variable $z = r^q$ confirms its uniform distribution.

In Fig. 4d the density (histogram) of variable $z = r^q$ confirms its uniform distribution.

Exponent q is called the *distribution mapping exponent* or – referring to the fractal nature of data – the *scaling exponent* as it scales distances according to the scaling law $z = r^q$.

We omit formal definitions here. They were published elsewhere [5], [6], [7].

4.3 The Correlation Integral as a Mean of Local Functions

Let us return to Eq. (1). In this section we show that the correlation integral is the mean of distribution mapping functions, as has already been shown in [11].
Let $h(x)$ be the Heaviside step function. Then the correlation integral is

$$C_I(r) = \lim_{N\to\infty} \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N-1} h(r - l_{ij})$$

and also

$$C_I(r) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{N-1} \sum_{j=1}^{N-1} h(r - l_{ij}) \right) .$$

Comparing the last formula with (2), we get directly

$$C_I(r) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} D(x_i, r) \quad . \qquad (4)$$

4.4 Log Linear Transformation

If the distribution mapping exponent is known, it holds another relation (see (3))

$$P(l < d) = Cd^q, \quad (5)$$

where $P(.)$ is the probability. This relation is linear with respect to variable

$$z = d^q . \quad (6)$$

Substituting (6) into (5), we get a linear relation and by derivation according to z, we get density of the random variable z

$$p = C.$$

This consideration is related to one fixed point x. For each point there is a particular value of the distribution mapping exponent q. But, as shown in [5], the value of q lies in rather a narrow band around its mean value for all points, and the correlation dimension v can be stated as this mean. The correlation dimension is the scaling exponent that holds for the whole data set. Taking the correlation dimension as a representative value for the distribution mapping exponent, the equation above can be rewritten with v instead of q, especially

$$P(l < d) = Cd^v, \quad (7)$$

and

$$z = d^v . \quad (8)$$

From the elementary theory of probability it follows that if a distribution function is linear with respect to random variable then the corresponding density is constant, i.e. $p(l^v < z) = C$ for z from (0, z_0).

Returning to (8), by logarithm we get the relation

$$\ln z = v \ln d, \quad (9)$$

that can also be written as

$$Z = vD, \quad (10)$$

where $Z = \ln z$ and $D = \ln d$.

Eq. (10) can be considered as linear expansion of logarithm of distance that leads finally to uniform distribution of scaled distances according to (7).

This simple finding leads to interesting conclusions.

First, after transformation (8), where $d < d_0$ all multidimensional data have uniform distribution (7).

Second, it means that due to measurement of the distance by z (instead of d) it all holds for any point of the data space and then data (from point of view of distances) is isotropic for $z < z_0$.

Third, polynomial transformation (expansion) (8), is a linear transformation of logarithms (9).

5 DISCUSSION

Comparing polynomial transformation (expansion) (8), especially in the form (10) with the Hubble law (1), one can - using the terminology of the expansion of the Universe - say that equation (10) describes isotropic expansion of data (naturally, measured in logarithmic scales).

The difference between Hubble's equation (1) and equation (10) is that (1) describes certain physical dynamics (at the left side is the zooming out speed), while (10) is a static logarithmic transformation of distances. Very important for physics and astronomy is that Hubble's law, i.e. receding of objects, holds for any point (observer) in the Universe and it does not matter in which direction the observer is looking.

Formal similarity of (1) and (10) shows the analogy of the correlation dimension with the Hubble constant. The correlation dimension is also a constant of proportionality, here in the logarithmic scale of distances of multidimensional data from any point x ("observer"). Due to this, the space of multidimensional data appears in logarithmic scale of distances as isotropic.

Finding that the Universe is isotropic, i.e. that it recedes everywhere, does not mean that objects are spread uniformly. Similarly, the transform (9), (10) represents isotropic, identical in all places transformation of logarithms of distances from point x, but it does not mean uniform distribution of points in the data space.

In summary, using polynomial expansion of distances according to (9) or (10) means to measure distances in rather strange units, e.g. meters to the q or correlation dimension v – power

instead of meters directly. The q and v are generally real numbers and they are lesser than the space dimension n. It can be seen that this transformation, the use of this strange measurement unit, causes that the concentration phenomenon is eliminated and the data space appears to be filled with data points so that the number of points in a ball is proportional to its radius measured in r^v, e.g. in meters to the correlation dimension power (!). Moreover, it does not depend on the point from which the distance is measured. From this point of view, the data space appears to be isotropic.

There arises the question what is this good for? The transformation introduced was used when designing classifiers that work rather well and are described in [5], [6], [7].

6 CONCLUSION

Using decomposition of the correlation integral to local functions, we have shown that particular functions are the distribution mapping functions that can be considered as distribution functions of distances of neighbors of a fixed point. It was also shown that there exists a simple polynomial approximation of this distribution function in the form $P(l < d) = Cd^q$, where q is a local scaling exponent that is very close to the correlation dimension v. Using variable $z = d^q$ or $z = d^v$, the distribution mapping function as a distribution function of random variable z appears to grow linearly with z and thus z has uniform distribution. In this way we have found that using polynomial expansion of distances from point x the data space appears to be uniform. The polynomial expansion of distances is nothing else than linear transformation of logarithms of distances with proportionality constant equal to the correlation dimension. This is valid for any data point x, i.e. any place in the data space and thus this phenomenon is isotropic. Analogy with the Hubble law is apparent – linear growth with logarithm of distance and isotropy.

7 ACKNOWLEDGEMENT

This work was supported by the Ministry of Education of the Czech Rep. under the INGO project No. LG 12020. Results presented here are included into the software package IINC used in the LG 12020 project.

8 REFERENCES

[1] M. Talagand, "A New Look at Independence", The Annals of Probability, Vol. 24, 1996, No.1, pp.1-34.

[2] F. Murtagh, "The Remarkable Simplicity of Very High Dimensional Data: Application of Model-Based Clustering", Journal of Classification Vol. 26, 2009, pp. 249-277.

[3] B. B. Mandelbrot, The Fractal Theory of Nature. W. H. Freeman and Co., New York, 1982, pp. 360-361.

[4] P. Grassberger, I. Procaccia, "Measuring the strangeness of strange attractors", Physica Vol. 9D, 1983, pp. 189-208.

[5] M. Jiřina, M. Jiřina, Jr., "Correlation Dimension-Based Classifer ", IEEE Transactions on Cybernetics, Vol. 44, 2014,in print.

[6] M. Jirina, M Jirina Jr., "Utilization of singularity exponent in nearest neighbor based classifier", Journal of Classification, Vol. 30, 2013, No. 1, pp. 3-29.

[7] M. Jiřina, M. Jiřina Jr., "Classification Using Zipfian Kernel", Journal of Classification (Springer), Vol. 31, 2013, in print.

[8] P. Camastra, A. Vinciarelli, "Intrinsic Dimension Estimation of Data: An Approach based on Grassberger-Procaccia's Algorithm", Neural Processing Letters Vol. 14, 2001, No. 1, pp. 27–34.

[9] F. Takens, "On the Numerical Determination of the Dimension of the Attractor". In: Dynamical Systems and Bifurcations. Lecture Notes in Mathematics, Springer, Berlin, Vol. 1125, 1985, pp. 99–106.

[10] F. Camastra, "Data dimensionality estimation methods: a survey", Pattern Recognition Vol. 6, 2003, pp. 2945–2954.

[11] M. Jiřina, M. Jiřina, Jr., „Classification by the Use of Decomposition of Correlation Integral." In: Abraham, A.; Hassanien, A.; Snášel, V. (Eds.) Foundations of Computational Intelligence (Studies in Computational Intelligence 205), Springer, Berlin, 2009, pp. 39-55.