International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 5 (2013) pp. 159-166

Near Neighbor Distribution in Sets of Fractal Nature

Marcel Jiřina
Institute of Computer Science AS CR
Dept. of Nonlinear Modeling
Pod Vodárenskou Věží 2
182 07 Prague, Czech Republic
e-mail: marcel@cs.cas.cz

***Abstract*: Distances of several nearest neighbors of a given point in a multidimensional space play an important role in some tasks of data mining. Here we analyze these distances as random variables defined to be functions of a given point and its *k*-th nearest neighbor. We prove that if there is a constant *q* such that the mean *k*-th neighbor distance to this constant power is proportional to the near neighbor index *k* then its distance to this constant power converges to the Erlang distribution of order *k*. We also show that constant *q* is the scaling exponent known from the theory of multifractals.**

Keywords: *nearest neighbor, fractal set, multifractal, Erlang distribution*

I. Introduction

There are distinct problems in dealing with the nearest neighbor or several nearest neighbors to a given point in data mining methods and procedures. A rather strange behavior of the nearest neighbors in high dimensional spaces was studied from different points of view.

One of them is a nonparametric technique of probability density estimation in multidimensional data space [1] , [2] . Classification into two or more classes and the probability density estimate using several nearest neighbors is a typical task [1] , [2] , [5] , [6] . Also clustering is a problem where interpoint distances are in common use. This class of problems looks for the highest quality of the probability density estimation, while efficiency and speed are secondary.

The other issue is the problem of nearest neighbor searching. This task is interesting and important in database applications [7] , [8] , [9] . A typical task of searching in large databases is searching for other nearest neighbor queries. For this class of problems, maximal performance, i.e. the speed of nearest neighbor searching, is a primary task.

The use of the *k*-nearest neighbor approach for other applications was studied, also e.g. by [10] , and [11] .

For the problem of searching the nearest neighbor in large databases, the boundary phenomenon was studied in [8] . It was shown by the use of L_{max} metric why the actual performances of the nearest neighbor searching algorithms tend to be much better than their theoretical analyses would suggest. The cause is just the boundary effect in a high dimensional space. In [9] it was found that as dimensionality increases, the distance to the nearest data point approaches the distance to the farthest data point of the learning set.

For probability density estimation by the *k*-nearest -neighbor method, the best value of *k* must be carefully tuned to find optimal results. The value of *k* is also dependent on the total number *N* of samples of the learning (or training) data set. Let a point *x* (a query point [7]) be given and let there be a ball with its center in *x* and containing *k* points of the learning set in it. Let the volume of this ball be V_k. Then, for the probability density estimate in point *x* it holds [1]

$$p_k(x) = \frac{k/N}{V_k}. \quad (1)$$

A simple experiment with density estimation at the center of a multidimensional cube with uniformly distributed points will show that starting from some *k* the value of $p_k(x)$ for larger *k* is not constant, as it should be, but lessens. The boundary effect surprisingly influences the *k*-th nearest neighbor about which everybody would say that this neighbor is still very near to point *x*.

In both cases, a statistics of interpoint distances is of primary interest. One often-used tool is the Ripley *K*-function [12] usually written in the form:

$K(r) = \lambda^{-1} E$(number of further points within distance *r* of an arbitrary point),

where λ is the intensity of the process or the expected number of events per unit of area (assumed constant). The *K*-function has been studied from different points of view in vast literature.

In some works, the point process is limited to two dimensions and *K*-function as well [13] , [14] . Marcon and Puech [14] summarize important features of the K-function including different corrections of edge (boundary) effects.
They summarize important features of the *K*-function including different corrections of edge (boundary) effects. An edge effect means that points located close to the domain borders are problematic because a part of the circle inside

which points are supposed to be counted is outside the domain. Result of ignoring this effect is underestimating K.

D. Evans [15] studies a class of random variables defined to be functions of a given point and its nearest neighbors. If the sample points are independent and identically distributed, the associated random variables will also be identically distributed, but not independent. He shows that random variables of this type satisfy a strong law of large numbers, in the sense that their sample means converge to their expected values almost surely as the number of sample points $n \to \infty$. Moreover, Evans introduced an interesting lemma that for every countable set of points in R^d, any point can be the (first) nearest neighbor of at most int(dV_{dp}) other points of the set, where V_{dp} is the volume of the unit ball in R^d with L_p norm.

Dealing with K-functions, it was already Ripley [12] who pointed out that the K-function shares some of the properties of the interpoint distribution function, even though it is not a distribution function because $K(r) \to \infty$ as $r \to \infty$. Later on, Bonetti and Pagano [16] introduced empirical cumulative distribution function $F_N(d)$ of distances d for total N samples. This is, in fact, a correlation integral [17] in one of its many forms. Bonetti and Pagano then gave a proof that the scaled distribution of $F_N(d)$ computed at a finite set of values, d_1, d_2, ... converges to a multivariate normal distribution as $N \to \infty$. The same result was given by a different way already by Silverman [18] .

The goal of this study is to analyze the distances of the nearest neighbors from given point x and the distances between two of these neighbors, the i-th and (i-1)-st in the space of randomly distributed points. We show that the k-th nearest neighbor distance from a fixed point (the query point) is distributed according to modified Erlang distribution of order k. The modification depends on that one uses a variable that is equal to the distance to the proper exponent power instead of the distance alone as an independent variable.

In this paper we first point out some features of neighbor distances in a multidimensional space. Second, we remind of the probability distribution mapping function, and the distribution density mapping function. These functions map probability distribution of points in R^n to a similar distribution in the space of distances, which is one-dimensional, i.e. R^1. Third, influence of boundary effects on the probability distribution mapping function is shown, and the power approximation of the probability distribution mapping function in the form of (distance)^q is introduced. We show that exponent q is, in fact, the scaling exponent known from the theory of multifractals [19] , [20] , [21] . Non-uniform as well as uniform distributions are considered. In conclusion, we can say that the nearest neighbor space does not look so strange as shown in [7] , [8] when we look at it from the point of view of a nonlinear scale measured by suitable power q of neighbors distances.

II. Near Neighbors Distribution Problem

The nearest-neighbor-based methods usually use (1) for a probability density estimate and are based on the distances of neighbors from a given point. Using the neighbor distances or interpoint distances for the probability density estimation should copy the features of the probability density function based on real data. The idea of most near-neighbors-based methods as well as kernel methods [2] assumes a reasonable statistical distribution in the neighborhood of the point in question. That means that for any point x (the query point [7]), the statistical distribution of the data points x_i surrounding it is supposed to be independent of the location of the neighbor points and their distances x_i from point x. This assumption is often not met, especially for small data sets and higher dimensions.

To illustrate this, let us consider Euclidean metrics and uniformly distributed points in cube $(-0.5, +0.5)^n$. Let there be a ball with its center in the origin and the radius equal to 0.5. This ball occupies $\frac{4}{3}\pi.0.5^3 = 0.524$, i.e. more than 52 % of that cube in a three-dimensional space, 0.080746, i.e. 8 % of unit cube in 6-dimensional space, 0.0026 in 10-dimensional space, and $3.28e^{-21}$ in 40-dimensional space. It is then seen that starting by some dimension n, say 5 or 6 and some index i, the i-th nearest neighbor does not lie in such a ball around point x but somewhere "in the corner" of that cube but outside this ball. Drawing a larger ball, we can see that some parts of it are empty, especially parts near to its surface. In farther places from the origin, the space thus seems to be less dense than near the origin. It follows that this i-th neighbor lies farther from point x as would follow from the supposed uniformity of distribution. The function $f(i) = \bar{r}_i^n$, where $\bar{r}_i$ is the mean distance of the i-th neighbor from point x, should grow linearly with index i in the case of uniform distribution without the boundary effect mentioned. In the other case, this function grows faster than linearly.

There is some unknown distribution of points in the neighborhood of point x. If this distribution is not uniform, we would like to have function $f(i)$ "uniform" in a sense that $f(i)$ is proportional to number i. The best choice would be $f(i) = \bar{r}_i^n$ for uniform distribution and no boundary effects. In real cases of higher dimensions, boundary effects occur every time. Not to neglect these effects, let us choose a function $f(i) = \bar{r}_i^q$, where q is a suitable power, $q \le n$. A suitable choice of q will be discussed later.

III. Fractal systems

One of the most important elements of the chaos theory are singularity exponents (also called scaling exponents). They are used in multifractal chaotic series analysis. We try here to use these exponents in a formula for probability distribution of near neighbors of a given position x. This task usually has nothing to do with time series but as shown already by Mandelbrot in 1982 [19] any data may possess a fractal or multifractal nature.

It can be found that the chaos theory provides some useful elements and tools that could be utilized for estimating the probability mentioned, and consequently to use them, e.g. for classification tasks or for tasks of nearest neighbors searching. The chaos theory is focused on chaotic processes that are described by time series. Therefore, the order of values of variables plays significant role. There is a lot of practical tasks described by data that do not form a series. In spite of that some elements of the chaos theory could be used for

processing of data of this kind [27] . Such data are, for example, the well-known iris data on three species of iris flower given by Fisher [33] . Each individual sample describes one particular flower but neither flowers nor data about them form series. There is a set of flowers as well as a set of data without any ordering. The task is to state to what species a flower belongs according to measured data. Also one can define a distance or dissimilarity defined on the parameter space that describes individual flowers as points. Now one can define a distance between two flowers and also for a particular flower to find the nearest neighbor, i.e. the most similar one, the second nearest, i.e. the second most similar flower and so on.

Fractal systems are known to exhibit a fractional power function dependence between a variable a called a *scale* (of a resolution quantity, or of a *measure* that can be associated with a nonuniform distribution with a support defined for the fractal set) and frequency s or of probability P of its appearance,

$$s(a) \approx a^h \quad \text{or} \quad P(a) \approx a^h .$$

Here h is called a fractal dimension of a fractal system.

A multifractal system is a generalization of a fractal system in which a single exponent h (the fractal dimension) is not enough to describe its behavior; instead, a continuous spectrum of exponents (the so-called singularity spectrum or Hausdorff spectrum) is needed. In a multifractal system, a local power law describes the behavior around any point x

$$s(x+a) - s(x) \approx a^{h(x)} . \qquad (2)$$

The exponent $h(\vec{x})$ is called the singularity (or holder) exponent or singularity strength [20] , as it describes the local degree of singularity or regularity around point $\vec{x}$, and a is called a scale of a multiresolution quantity or of a measure.

The ensemble formed by all the points that share the same singularity exponent is called the singularity manifold of exponent h, and is a fractal set of fractal dimension $D(h)$. The curve $D(h)$ vs. h is called the singularity spectrum or Hausdorff spectrum and fully describes the (statistical) distribution of variable a. The singularity spectrum of a monofractal consists of a single point.

For multifractal objects, one usually observes a global power law scaling of the form (2) or simply $P(a) \approx a^h$. That is a local power law that describes the behavior around any point $\vec{x}$ at least in some range of scales and for some range of h. When such a behavior is observed, one talks of scale invariance, self-similarity or multiscaling.

In a slightly different perspective, fractal appears as (innumerably many) points in a multidimensional space of dimension n. In practice, the fractal set is given, "sampled", by finite collection of M samples (patterns, measurements etc.) that form isolated points in n-dimensional space. In case of independent samples there is no ordering or sequence of samples, thus there is no trajectory. Samples can only be more similar or dissimilar in some sense. If the n-dimensional space is a metric space, the most common dissimilarity measure is a distance. For a collection of samples of the fractal set the mutual distances between samples (points) are different realizations of a random variable called distance a between samples. This random variable has support $(0, \infty)$. The distance a between samples has distribution function $F(a)$ and corresponding probability density $P(a)$.

Thus, we have introduced a fractal set, the support on it and multifractal measure $P(a)$. Indeed, having a pair No. i of two samples lying in distance a_i one from the other, then $P(a_i)$ is a probability of appearance a_i., i.e. of distance a_i at place i. Apparently, the way of enumerating can be arbitrary.

To state an empirical distribution $F(a_i)$ and density $P(a_i)$ let us enumerate (sort) all distances a_i, $i = 1, 2, .. M$, $M = \frac{1}{2}N(N-1)$ so that $a_i < a_{i+1}$. The distribution function $F(a_i)$ is then a staircase function of N steps, and $F(a_i) = i/M$. An empirical derivative of $F(a)$ at a_i is $P(a_i)$ and can be approximated as $P(a_1) = 0$, $P(a_i) = 1/(a_i - a_{i-1})$. More sophisticated approximations of $P(a_i)$ are possible. After that one can use formulas for Hausdorff dimension $f(q)$ and singularity strength $\alpha(q)$ above and get the singularity spectrum.

IV. Analysis

A. *Point process of neighbors*

Let us assume random distribution of points x_i, $i = 1, 2, \ldots, N$ in bounded region W of n-dimensional space R^n with L_p metrics. We consider appearance of neighbors around a given location (query point) x as a point process $\mathbf{P}$ in R^n [34] , [35] .

Throughout this part, we say that point x is inside S in the following sense: For each neighbor y considered, the ball with its center at x and radius equal to $\|x-y\|$ lies inside S. This is the case where the boundary effects do not take place.

B. *One-dimensional case*

In this particular case, we are interested in homogenous Poisson process only. Positions of points on R^+ from location $u \equiv 0$ are characterized by distances from point 0; the k-th neighbor of point 0 appears at distance r_k. We use simple analogy between time and distance here. In the one-dimensional homogenous Poisson process with intensity λ the inter-arrival times have an exponential distribution [34] . Then the distance Δ between two neighbor points is a random variable with exponential distribution function $P(\Delta) = 1 - e^{-\lambda\Delta}$ and probability density function $p(\Delta) = e^{-\lambda\Delta}$ [34] , [35] . For this distribution function the mean is $E\{\Delta\} = 1/\lambda = d$ and it is the mean distance between two neighbor points.

Let us imagine a positive half-line with a query point at point 0 and with randomly and uniformly spread points, i.e. the distance between two neighbor points is Δ with mean d. The question is: What is the distribution of the distance of each point from point 0? The distance of i-th point (i-th neighbor of point 0) is simply the sum of individual distances between two successive points. These individual distances are independent and have the same exponential distribution with $\lambda = 1/d$.

Because the distance is the sum of all successive distances, it is also the sum of independent exponentially distributed random variables. This problem was studied in connection with mass service (queuing) systems and represents the issue of i independent exponential servers working in series

(cascade) [22]. The servers have the same exponential distribution of the service time with constant λ. Generally, the resulting total service time after the i-th server is given by gamma distribution with integer first parameter or the Erlang distribution Erl(i, λ) [22] [23] , [24]

$$p_i(x) = \frac{1}{i!}\lambda^i x^{i-1} e^{-\lambda x} .$$

C. Multidimensional case

Let us have a spatial point process **P** in R^n. Point process **P** considered, as a set of points can be a fractal set. It is possible to define several kinds of distances between points of **P** [34] .

1. One can introduce a distance between two points of **P**, R_{ij}=dist(x_i; x_j), x_i, $xj \in$ **P**. In a bounded region $W \subset R^n$ a cumulative distribution function of R_{ij}

$$C_I(r) = \lim_{N\to\infty} \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N-1} h(r - l_{ij}) \qquad (3)$$

is denoted as correlation integral. $h(.)$ is the Heaviside step function. Grassberger and Procaccia [17] have introduced correlation dimension ν as limit

$$\nu = \lim_{r\to 0} \frac{C_I(r)}{r} \qquad (4)$$

2. Let R=dist(x, **P**) be the shortest distance from the given location u to the nearest point of **P**. This is called the contact distance and the cumulative distribution function of R=dist(x, **P**) in a bounded region $W \subset R^n$ is called contact distribution function or the empty space function [35] . We can extend distances from a given location u to the second, third,... nearest neighbor point of **P** and find the corresponding distribution function of reaching k-th nearest neighbor. This function is called a survivor function [36] .
3. For distances from a given (and fixed) location u to other points of **P** we use notion of a distribution mapping function (DMF) introduced in our previous works [25] , [26] .

Definition

The distribution density mapping function $d(x, r)$ of the neighborhood of the query point x is function

$$d(x,r) = \frac{\partial}{\partial r} D(x,r),$$

where $D(x, r)$ is a probability distribution mapping function of the query point x and radius r.

In bounded region W when using a proper normalization the DMF is, in fact, a cumulative distribution function of distances from a given location u to all other points of **P** in W. We call it also the near neighbors distribution function of point u. Here we use some definitions introduced in our previous works [25] , [26] .

Definition

The probability distribution mapping function $D(x, r)$ of the neighborhood of the query point x is function

$$D(x,r) = \int_{B(x,r)} p(z)dz ,$$

where r is the distance from the query point and $B(x, r)$ is the ball with center x and radius r.

Note. It is seen that for fixed x the function $D(x, r)$, $r > 0$ is monotonously growing from zero to one. Functions $D(x, r)$ and $d(x, r)$ for fixed x are the (cumulative) probability distribution function and the probability density function of the near neighbor distance r from point x, respectively. We use $D(x, r)$ and $d(x, r)$ mostly in this sense.

Definition

Let a, b be distances of two points from a query point x in R_n. Then

$$d_{(n)} = \left| a^n - b^n \right| .$$

We differentiate between the $d_{(n)}$ and distance $d = |a - b|$; we write also $d_{(n)}(a, b)$.

D. Scaling

Definition

Let there be a positive q such that

$$\frac{D(x,r)}{r^q} \to const \text{ for } r \to 0+ .$$

We call function $d_{(q)} = r^q$ a power approximation of the probability distribution mapping function.

This definition naturally follows a previous definition. The important thing is that the last definition exactly gives a true picture of the scaling property known from the fractal and multifractal systems theory, where q is known as *multifractal dimension, scaling* (*singularity* or *Hölder*) *exponent* or *singularity strength* [19] , [20] , [27] , [30] . If **P** is nonfractal, then scaling exponent q has integer value for all points of **P**. Especially if **P** is a homogenous Poisson process, then $q = n$ [36] . If **P** is a monofractal, scaling exponent q has the same value for all points of **P**, i.e. a local scaling exponent q is equal to correlation dimension ν. In the other case, **P** is a multifractal characterized by local scaling exponent $q(u)$ that depends on particular position of location (query point) u in question.

Moreover, there is a well-known correlation integral in form (3). The correlation integral can be rewritten in the form:

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

and also

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

Thus, the correlation integral is a mean of probability distribution mapping functions for all points of a set considered.

The estimated value of the correlation dimension computed using well-known log-log plot [17] depends heavily on the number of data points considered. This appears especially in high dimensional spaces. It has been discussed to a large extent in [29] , [30] , [31] . Krakovska [31] gave estimate of the error in correlation dimension estimate. Her estimate evaluates the influence of the edge (boundary) phenomenon for data in a cube. The cube was considered as the worst case as data usually form a more rounded "cloud" in the space. At the same time, she generated uniformly distributed data in a cube so that each coordinate was a random variable with uniform distribution on (0, 1). Thus, point process that generated points in multidimensional hypercube in this way was not a multidimensional Poisson process. Multivariate data generated by the multidimensional Poisson process have fractal dimension equal to the embedding space dimensionality n. Data generated by the method described above then apparently have lower fractal dimension and it can be easily seen that this is not caused by lack of data points or by edge effects (that can be eliminated by the use of the approach described in Chap. IV.A). We show it in greater detail in Chap. VI.

V. Results

A. *Uniform-like distribution and scaling exponent*

Let a query point x be surrounded by other points uniformly distributed to some distance d_0 from point x. It means that there is a ball $B(x, d_0)$ with points uniformly distributed inside it. In this case, the number of points in a ball neighborhood with center at the query point x grows with the n-th power of distance from the query point (up to distance d_0). The distribution mapping function $D(x, r)$ also grows with the n-th power of distance from the query point x, i.e. the number of points in ball with center at x and radius r grows linearly with r^n.

$$D(x, r) = (r / d_0)^n \ \text{ for } r \in \langle 0, d_0 \rangle,$$

and

$$D(x, r) = 1 \text{ for } r/d_0 > 1.$$

At the same time, $D(x, r)$, as a function of $z = r^n$, $D(x, z)$, grows linearly, too. The distribution density mapping function $d(x, z)$, taken as $\frac{\partial}{\partial(r^n)} D(x, r^n)$, is a *constant* for $r \in (0, d_0)$ and zero otherwise.

Let r_i be a distance of the i-th neighbor of point x, and $z_i = r_i^n$. It follows that the mean $d_{(n)}$, $\bar{d}_{(n)} = E\left(\left| r_{k-1}^n - r_k^n \right|\right)$ (point k-1 = 0 is point x) of successive neighbor points is a constant under the condition of uniform distribution.

Measuring "distance in n dimensions" by $(distance)^n$, i.e. by the use of $d_{(n)}$ we get, in fact, the same picture as in one-dimensional case, discussed in Sec. IV.B. Because the distribution of $d_{(n)}(0, r)$, $r \in \langle 0, d_0 \rangle$ is uniform, then $d_{(n)}$ of successive neighbors, $d_{(n)}(r_{k-1}, r_k)$ is a random variable with exponential distribution function. It also follows that the $d_{(n)}$ of the i-th nearest neighbor from the query point is given by the sum of $d_{(n)}$´s between the successive neighbors. Then, it is a random variable with the Erlang distribution Erl(i, λ), λ = $1/\bar{d}_{(n)}$, where $\bar{d}_{(n)}$ is mean $d_{(n)}$ between the successive neighbors.

Definition

Let there be ball $B(x,r)$ with center x and radius r. Let there be a positive constant λ. We say that points in ball $B(x,r)$ are spread uniformly with respect to $d_{(q)}$, if $d_{(q)}(r_{k-1}, r_k)$ of two successive neighbors is a random variable with exponential distribution function with parameter λ.

Note that the wording "... are spread uniformly with respect to $d_{(q)}$" comprises fractal nature of data as well as the edge (boundary) phenomenon. In the first case, the q is the local scaling exponent; in the other case, the q captures the boundary effect. Usually the value of q is given by a mix of these two phenomena and is lower than it would be for these two phenomena separately.

Theorem

Let, for query point $x \in R_n$, there exist a scaling exponent q and a distance r such that in ball $B(x,r)$ with center x and radius r the points are spread uniformly with respect to $d_{(q)}$. Let $d_{(q)}(x_i, x_{i+1})$ between two successive near neighbors of point x have mean $\delta = E(d_{(q)})$ and let $\lambda = 1/\delta$. Then, the $d_{(q)}$ of the k-th nearest neighbor of point x is the random variable with the Erlang distribution Erl($d_{(q)}$, k, λ), i.e.

$$F(d_{(q)}) = 1 - \exp(-\lambda d_{(q)}) \sum_{j=0}^{k-1} \frac{(\lambda d_{(q)})^j}{j!}$$

$$f(d_{(q)}) = \frac{\lambda^k}{k!} d_{(q)}^{k-1} \exp(-\lambda d_{(q)}) \ .$$

Proof. Let us denote the i-th nearest neighbor of the point x by x_i, its distance from point x by d_i, its $d_{(q)}(x, x_i)$ by $d_{(q)i}$. Let us introduce a mapping Z: $x_i \to R_1+$: $Z(x_i) = d_{(q)i}$, i.e. points x_i are mapped to points on the straight line, the query point x to point 0. Let total number of points x_i in ball $B(x,r)$ be N. Then, the distribution mapping function is $d(x, r) = const \frac{N}{r^q}$. It follows from the assumption that number of points at distance ρ from point x grows linearly with ρ^q . Then, there is a proportionality constant $\lambda = N/r^q$. It follows that in mapping Z

points x_1, x_2, ... are distributed randomly and uniformly, and mean $d_{(q)}$ of the neighbor pairs is $r^q / N = 1/\lambda$. We, then, have uniform distribution of points on $d_{(q)}$ and then the $d_{(q)}$ between the neighbor points has exponential distribution [23] with parameter λ. From it $d_{(q)}$ of the k-th point x_k from point x is given by the sum of $d_{(q)}$s between successive neighbors. Then, it is a random variable equal to the sum of random variables with identical exponential distribution [23] , [24] with parameter λ, then

$$d_{(q)}(x_k) = Erl(d_{(q)}, k, \lambda) \,.\, \square$$

Note that this theorem is a generalization of the theorem [27] derived with the use of the spatial point process. In [28] author speaks about generalized gamma distribution, but, at the same time, considers the shape parameter as a positive integer that is the special case often called the Erlang distribution. Moreover, he considers a volume of a ball in the embedding space of dimension m (positive integer; here denoted by n) while we use more general distribution mapping (scaling) exponent q (real, positive).

B. Empirical distribution and empirical DME

We have pointed out that the distribution mapping exponent is nothing else than the *multifractal dimension (also known as scaling* (*singularity* or *Hölder*) *exponent* or *singularity strength)* [17] , [19] , [20] , [21] , [27] . There is also close relation to a well-known correlation integral [17] and correlation dimension. We have shown that correlation integral can be understood as a mean of probability distribution mapping functions for all points of the data set.

The multifractal dimension, scaling (singularity or Hölder) exponent or singularity strength is often used for characterization of one dimensional or two-dimensional data, i.e. for signals and pictures. Our results are valid for multivariate data that need not form a series because data are considered as individual points in a multivariate space with proper metrics.

VI. Simulation analysis

The target of simulation is to demonstrate

1. That the distribution mapping exponent one can use as a relatively global feature for characterizing the data set given.
2. That the empirical distribution of variable r_q of nearest neighbors is really close to the Erlang distribution, as stated in Theorem 1.

For all simulations, 32 000 samples in n-dimensional cube were used. Data in a cube were generated so that each coordinate was a random variable with uniform distribution on (0, 1). Thus, the generated points in multidimensional hypercube were not points arising from a multidimensional Poisson process. Only multivariate data generated by the multidimensional Poisson process have fractal dimension equal to the embedding space dimensionality n. Data generated by method described then apparently have lower fractal dimension and it can be easily seen that this is not caused by lack of data points or by edge effects, as shown in the following.

After samples were generated, each sample was taken as a query point x and for each sample 10 nearest neighbors were found and their distances r_i, i = 1, 2, ... 10 recorded. After that a distribution-mapping exponent q was found as a value for which mean values of r_i^q grow linearly, i.e. successive differences of their means $d_{qi} = E(r_i^q) - E(r_{i-1}^q)$ are approximately constant; $r_0^q = 0$. In this experiment, results are apparently influenced by procedure of points generation because in an opposite case the distribution mapping exponent would be equal to data space dimension n. Truly, we organize the simulations this way to show points 1 and 2 above.

The values of q found for dimension n from 1 to 40 are shown in the second column of Table 1.

n	q found by simulation
1	1
5	4.8
10	9
20	15.6
40	28

Table 1. Values of the distribution-mapping exponent q for some uniform n-dimensional cubes found by simulation and approximated.

In Fig. 1, differences $d_{qi} = E(r_i^q) - E(r_{i-1}^q), r_0^q = 0$ for n = 20 and three different exponents are shown. Here it is seen that the value q = 15.6 gives approximately a horizontal straight line. It means that differences $d_{(q)i}$ are all nearly the same. This fact corresponds to approximately uniform distribution of variable r^q. Then, q = 15.6 is a good estimate for the distribution-mapping exponent in this case.

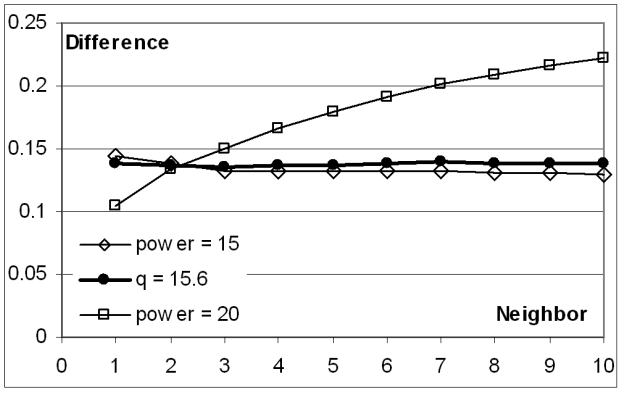

Figure 1. Differences $d_{qi} = E(r_i^q) - E(r_{i-1}^q), r_0^q = 0$ for n = 20, 10 nearest neighbors, and three different exponents.

In Fig. 2, histograms of $d_{(q)i} = r_i^q$ for q = 15.6, n = 20 and for the first 10 neighbors (i = 1, 2, ... 10) are shown. The histograms were smoothed using averaging over five values. The bin size is 0.01. The histograms show, in fact, probability

density functions for the Erlang distribution for indexes 1 (exponential) to 10. This is just what was expected when the probability distribution mapping function with a proper distribution-mapping exponent was used.

VII. Discussion

Here we discuss a problem of empirical value of the distribution mapping exponent and its use.

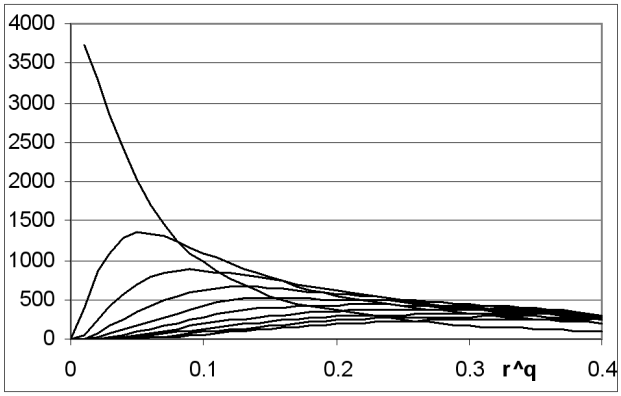

Figure 2. Histograms of variable $z = r_i^q$, $q = 15.6$, $n = 20$. Lines from top to bottom $i = 1, 2, \ldots 10$.

In statistics, it is common to differentiate theoretical and empirical distribution. As to scaling exponent, no such a notion was introduced in the literature about fractals and multifractals. Discrepancies between theoretical value and empirically stated value are explained by lack of data, and either more data are demanded or some corrections are made. Such discrepancies and errors were widely discussed, see e.g. [29] , [30] , [31] . In [31] correction factors for different embedding dimensions are presented, based on the assumption of data uniformly distributed in a cube. There is a general objection that data do not form a cube and that cube is, then, too strict a model for edge phenomenon. Moreover, there is a slightly problematic way of how "uniform" data are generated, as discussed above.

Here we do not use such corrections but an estimated empirical value. The empirical value is influenced first by true local scaling exponent of the data generating process, and second by edge effect caused by limited amount of data. Corrections discussed above try to eliminate this influence. In the case of probability density estimation and classification and in other tasks dealing with neighbor's distances one needs to work with the neighbor's distance distribution as it appears in a given empirical environment, and idealized limit case of the number of data points going to infinity is found as impractical. Instead, a large number of experiments with the same finite amount of data are considered; with the number of such experiments eventually going to infinity.

Now consider a distribution of k-th nearest neighbor. Let us use the coefficient of variation C_V that is a normalized measure of dispersion of a probability distribution. The coefficient of variation C_V is defined as the ratio of the standard deviation to the mean. For the Erlang distribution of k-th order it (results in an?) interesting formula for the coefficient of variation $C_V = \sigma/\mu = 1/\sqrt{k}$. This relation shows that relative spread of variable with the Erlang distribution of the k-th order diminishes to zero with order k going to infinity. On the other hand, for $k = 1$ (in fact exponential distribution), i.e. for the nearest neighbor, the $C_{V1} = 1$, while for the second nearest neighbor ($k = 2$) there is $C_{V2} = \sqrt{2}/2 \approx 0.707$. This shows why in [25] , [26] we do not recommend to use the first nearest neighbor of each class in contrast to the old finding by Cover and Hart [32] that the first nearest neighbor brings half of information about class of the query point.

Because the Erlang distribution converges to Gaussian distribution for order $k \to \infty$, the result according to Theorem 1 also contains some results of e.g. [15] , [16] , [18] , [28] about convergence of near-neighbor distances.

VIII. Conclusion

When using the notion of distance, there is a loss of information on the true distribution of points in the neighborhood of the query point. It is known [7] , [8] that for larger dimensions something like local approximation of real distribution by uniform distribution in practice does not exist. We have also shown why. On the other hand, the assumption of at least local uniformity in the neighborhood of a query point is usually inherent in the methods based on the distances of neighbors.

Introducing power approximation of a distribution mapping function here solves this problem. It has been shown that the exponent of the power approximation is the scaling exponent known from the theory of multifractals for the number of points going to infinity. For finite set, this exponent includes boundary effects. By using the scaling exponent, the real empirical distribution is transformed to appear as locally uniform. It follows that when using exponentially scaled distance of the k-th neighbor the scaled distance has the Erlang distribution of order k.

Acknowledgment

This work was supported by Ministry of Education of the Czech Republic under INGO project No. LG 12020.

References

[1] Duda, R.O., Hart, P.E., Stork, D.G., *Pattern Classification. Second Edition.* John Wiley and Sons, Inc., New York, 2000.

[2] Silverman, B. W., *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

[3] Qamar, A.M., Gaussier, E., RELIEF Algorithm and Similarity Learning for k-NN. *International Journal of Computer Information Systems and Industrial Management Applications,* Vol 4 (2012), pp. 445-458. Available at

[4] http://www.mirlabs.org/ijcisim/regular_papers_2012/Paper49.pdf

[5] Eadie, Wt. T. et al., *Statistical Methods in Experimental Physics.* North-Holland, 1982.

[6] Moore, D.S., Yackel, J.W.: Consistency Properties of Nearest Neighbor Density Function Estimators. *Annals of Statistics,* Vol. 5, No. 1, pp. 143-154 (1977)
[7] Hinnenburg, A., Aggarwal, C.C., Keim, D.A., What is the nearest neighbor in high dimensional spaces? *Proc. of the 26th VLDB Conf.,* Cairo, Egypt, 2000, pp. 506-515.
[8] Arya, S., Mount, D.M., Narayan, O., Accounting for Boundary Effects in Nearest Neighbor Searching. *Discrete and Computational Geometry,* Vol. 16 (1996), pp. 155-176.
[9] Beyer, K. et al., When is "Nearest Neighbor" Meaningful? *Proc. of the 7th International Conference on Database Theory.* Jerusalem, Israel, 10-12 January 1999, pp. 217-235.
[10] Biau, G., Cadre, B., Rouviere, L.: Statistical Analysis of k-nearest Neighbor Collaborative Recommendation. *The Annals of Statistics,* Vol. 38, No. 3, pp. 1568–1592, (2010)
[11] Stute, W.: Asymptotic Normality of Nearest Neighbor Regression Function Estimates. *The Annals of Statistics,* Vol. 12, No. 3., pp. 917-926, (1984)
[12] Ripley, B.D.: Modelling Spatial Patterns. *Journal of the Royal Statistical Society. Series B (Methodological),* Vol. 39, No. 2, pp. 172-212 (1977)
[13] Dixon, P.M.: Ripley's K function. In: Abdel H. El-Shaarawi and Walter W. Piegorsch (Eds.): *Encyclopedia of Environmetrics, Volume 3,* pp. 1796–1803 (ISBN 0471 899976), John Wiley & Sons, Ltd, Chichester, (2002).
[14] Marcon, E., Puech, F.: Evaluating the Geographic Concentration of Industries Using Distance-based Methods. *Journal of Economic Geography,* Vol. 3, No. 4, pp. 409–428 (2003)
[15] Evans, D.: A law of large numbers for nearest neighbor statistics. *Proc. R. Soc. A.,* Vol. 464, pp. 3175–3192 (2008)
[16] Bonetti M, Pagano M.: The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. *Stat. Med.* Vol. 24, No. 5, pp. 753-773 (2005)
[17] Grassberger, P., Procaccia, I.. Measuring the strangeness of strange attractors, *Physica,* Vol. 9D, pp. 189-208, (1983).
[18] Silverman, B.W.: Limit theorems for dissociated random variables. *Advances in Applied Probability,* Vol. 8, pp. 806-819 (1976)
[19] Mandelbrot, B.B.: *The Fractal Geometry of Nature,* W. H. Freeman & Co; ISBN 0-7167-1186-9 (1982).
[20] Chhabra, A., Jensen, R.V.: Direct Determination of the f(α) Singularity Spectrum. *Physical Review Letters,* Vol. 62, No. 12, pp. 1327-1330 (1989)
[21] Hu, J., Gao, J. and Wang, X.: Multifractal analysis of sunspot time series: the effects of the 11-year cycle and Fourier truncation. *Journal of Statistical Mechanics* P2066, 20 pp. (2009)
[22] Trivedi, K.S.: *Probability and Statistics with Reliability, Queuing, and Computer Science Applications.* Prentice-Hall, Inc., Englewood Cliffs, NJ 07632, USA, (1982)
[23] Kleinrock, L., *Queueing Systems, Volume I: Theory.* John Wiley & Sons, New York, 1975.
[24] Demaret, J.C., Gareet, A., Sum of Exponential Random Variables. *AEÜ,* Vol. 31, 1977, No. 11, pp. 445-448.
[25] Jiřina, M. Classification of Multivariate Data Using Distribution Mapping Exponent. In: *Proceedings of the International Conference in Memoriam John von Neumann,* pp. 155-168. Budapest: Budapest Muszaki Foiskola (Budapest Polytechnic), ISBN 963-7154-21-3. [International Conference in Memoriam John von Neumann, Budapest, 12.12.2003, HU] (2003)
[26] Jiřina, M. Local Estimate of Distribution Mapping Exponent for Classification of Multivariate Data. In: *Engineering of Intelligent Systems.* Millet : ICSC, 2004. s. 1-6. ISBN 3-906454-35-5. [International ICSC Symposium on Engineering of Intelligent Systems /4./, Madeira, 29.02.2004-02.03.2004, PT] (2004)
[27] Mandelbrot, B.B., Calvet, L., Fisher, A., 1997. A Multifractal Model of Asset Returns. *Working Paper. Yale University. Cowles Foundation Discussion Paper* #1164. Available from: http://ideas.uqam.ca/ideas/data/Papers /cwlcwldpp1164.html.
[28] Haenggi, M.: On Distances in Uniformly Random Networks. *IEEE Trans on Information Theory,* Vol. 51, No. 10, Oct. 2005, pp. 3584-3586.
[29] Smith, L.A.: Intrinsic limits on dimension calculations. *Physics Letters A,* Vol. 133, No. 6, pp. 283-288 (1988)
[30] Theiler, J.: Estimating fractal dimension. *J. Opt. Soc. Am. A,* Vol. 7, No. 6, pp. 1055-1073. (1990)
[31] Krakovská, A.: Correlation dimension underestimation. *Acta Physica Slovaca,* Vol. 45, No. 5, pp. 567-574 (1995)
[32] Cover, T.M., Hart, P.E.: Nearest Neighbor Pattern Classification. *IEEE Trans. on Information Theory,* Vol. IT-13, No. 1, pp. 21-27 (1967)
[33] Fisher,R.A.: The use of multiple measurements in taxonomic problems. *Annual Eugenics,* Vol. 7, Part II, pp. 179-188 (1936); also in: "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).
[34] Diggle, P.J.: *Statistical Analysis of Spatial Point Patterns.* ARNOLD, London, 2003.
[35] Baddeley, A: Spatial point processes and their applications. *Lecture Notes in Mathematics* 1892, Springer-Verlag, Berlin, 2007, pp. 1–75.
[36] Daley, D. J., Vere-Jones, D. *An Introduction to the Theory of Point Processes.Vol. I: Elementary theory and methods.* Springer, 2005.

Author Biography

Marcel Jiřina Education & Work experience: Faculty of Electrical Engineering, CTU Prague (1962), Faculty of Mathematics and Physics, Charles University Prague (1974), Ph.D. (1972). Head of Department of System Design at the Research Institute of Mathematical Machines Prague (1962-1986), researcher at the Institute of Computer Applications in Management, Prague (1986-1989), currently researcher at the Institute of Computer Science of Academy of Sciences CR, at the Department of Nonlinear Modeling.