RUDN Journal of MIPh
Вестник РУДН. Серия МИФ
2018 Vol. 26 No. 1 58–73
http://journals.rudn.ru/miph

UDC 004.021, 004.8, 519.6
DOI: 10.22363/2312-9735-2018-26-1-58-73

On a Method of Multivariate Density Estimate Based on Nearest Neighbours Graphs

Gleb Beliakov

School of Information Technology
Deakin University
221 Burwood Hwy, Burwood 3125, Australia

A method of multivariate density estimation based on the reweighted nearest neighbours, mimicking the natural neighbours techniques, is presented. Estimation of multivariate density is important for machine learning, astronomy, biology, physics and econometrics. A 2-additive fuzzy measure is constructed based on proxies for pairwise interaction indices. The neighbours of a point lying in nearly the same direction are treated as redundant, and the contribution of the farthest neighbour is transferred to the nearer neighbour. The calculation of the local point density estimate is performed by the discrete Choquet integral, so that the contributions of the neighbours all around that point are accounted for. This way an approximation to the Sibson's natural neighbours is computed. The method relieves the computational burden of the Delaunay tessellation-based natural neighbours approach in higher dimensions, whose complexity is exponential in the dimension of the data. This method is suitable for density estimates of structured data (possibly lying on lower dimensional manifolds), as the nearest neighbours differ significantly from the natural neighbours in this case.

Key words and phrases: density estimate, nearest neighbours, Choquet integral, fuzzy measure, natural neighbours

1. Introduction and Problem Formulation

Multivariate density estimates from finite samples play an important role in data analysis and clustering [1]. Among other applications, density estimates provide a way to construct density based metrics [2], density based averages [3, 4], perform density based clustering, and also compute robustly the mode(s) of a distribution [5, 6]. Practical applications include data analysis and machine learning, anomaly detection, econometrics, high energy physics, astronomy, flow cytometry, image analysis and computer vision to name a few. For example, spatial distribution of cosmic matter at megaparsec scale was analysed by using nonparametric density estimates in [7]. Density based metrics are often used in unsupervised data analysis, e.g., in the DBSCAN algorithm [8].

Histograms are traditionally used as density estimates of single variable distributions. Their use in the multivariate setting is problematic because of the rapidly growing number of histogram bins, the majority of which remain empty. Kernel-based density estimates due to the works by Parzen and Rosenblatt [1,5], often called Parzen-Rosenblatt windows, is a popular multivariate approach, in which a point density estimate is constructed by averaging the values of a kernel function of the distances between a fixed point and the data. One problem with kernel density estimates is the bandwidth selection, which is the smoothing parameter in this process. The values of the bandwidth parameter which are too small result in spiky estimates, values that are too large result in oversmoothing. There are approaches for automatic bandwidth selection based on cross-validation [9] but they are computationally expensive.

Another family of density estimates is based on the notion of the Voronoi diagram [10]. A Voronoi cell is a set of points which are closer to one point from the sample than to any other point in that sample. Intuitively, the volumes of Voronoi cells can serve as proxies

Received 5th October, 2017.

for density estimation: small Voronoi cells imply high density. One can view Voronoi cells as (polyhedral) bins in a histogram that contain a single datum. From the technical viewpoint, Voronoi cells are not very convenient, as a) there are Voronoi cells of infinite volume, and b) multiple calculation of Voronoi cell volumes is computationally expensive. Instead the dual of the Voronoi diagram, the Delaunay tessellation, is used [11]. The Delaunay tessellation is a partition of the convex hull of the data (and hence Delaunay cells are finite), and since these cells are simplices, their volume is computed easily in the multivariate setting. The method in [7] averages the reciprocals of the volumes of the neighbouring Delaunay simplices to provide point density estimates at every point in the sample. One important feature of Delaunay tessellation is that the neighbouring simplices involve data located all around the point at which the density estimate is computed. This feature led to the development of the method of "natural neighbours" in scattered data interpolation [12].

The issue with Delaunay tessellation is its complexity: the number of Delaunay cells grows exponentially with the dimension d of the space, more precisely as $O\left(n^{\lceil d/2 \rceil}\right)$, where n is the sample size. This is a manifestation of the course of dimensionality.

Another approach to density estimation is based on the nearest neighbour type graphs, including the k nearest neighbour graph (kNN), minimum spanning tree (MST) and Gabriel graph [13, 14]. The distance from a point to its nearest neighbour can give an estimate of the density, as it provides the volume of an empty sphere near that point. Compared to the Delaunay tessellation, there is no combinatorial explosion of complexity with the increasing dimension, as no space partitioning is required (only n^2 pairwise distances are needed to construct the MST or kNN graph). The MST and Gabriel graphs are subgraphs of the Delaunay graph, which prompted their use as proxies for the Delaunay tessellation. But on the other hand, the nearest neighbours are not always located all around a query point, and the nearest neighbour relation is not reciprocal. The kNN graphs may not be connected, which makes them not fully suitable for proximity calculations [14]. Selecting a larger value of k also leads to oversmoothing.

In this paper we explore one method of density estimation based on the nearest neighbours graph. In this method we take a sufficiently large value of k in the kNN density estimate, but ensure that only the neighbours located *all around* a query point are counted. That is, we attempt to marry the kNN with the natural neighbours approach, but without performing expensive Delaunay tessellation. To this end we use the notion of the discrete Choquet integral with respect to a specially constructed fuzzy measure. It allows one to account for correlations between the inputs, and explicitly model the notions of redundancy and positive reinforcement. In particular we account for contributions of the neighbours situated in the same direction from a query point and downweight the contribution of the furthest. This way only the contributions to the density estimate from the neighbours all around a query point will count.

The problem is formulated as follows. Given a sample of (independent, identically distributed) data of size n and dimension d,

$$\mathcal{D} = \left\{x_i^j\right\}_{i=1,\ldots,d;\ j=1,\ldots,n} = \left\{\left(x_1^j, \ldots, x_d^j\right)\right\}, \quad j = 1, \ldots, n,$$

find a density estimate approximating the probability density of the distribution the data were drawn from.

The paper is structured as follows. Section 2 presents the background material needed for the rest of the paper. Section 3 presents the proposed kNN reweighting method, including the construction of a 2-additive fuzzy measure from the interaction indices and computation of the threshold for the size of the cone of directions in the multivariate setting, so that the proportion of data located in such a code remains constant in different dimensions. Section 4 provides a numerical illustration and Section 5 concludes.

2. Preliminaries

2.1. Point Density Estimation Problem

Let the data set $\mathcal{D}$ be generated by sampling from a distribution with probability density $\rho : \mathbb{R}^d \to [0, 1]$. The goal of density estimation is to recover an approximation to ρ, denoted $\hat{\rho}$. Non-parametric methods do not assume any specific form of ρ and hence build $\hat{\rho}$ based only on the data.

Building a histogram is the traditional approach which usually works in one or two dimensions, but is not suitable in the multivariate setting because of the rapidly growing number of histogram bins where the data are allocated. There are several approaches to density estimation mentioned in the Introduction. In particular, kernel density estimates provide density $\hat{\rho}(\mathbf{x})$ at a point using that point as a centre of a neighbourhood of selected radius, while Voronoi diagrams provide point density estimates using the nearest neighbours of the point $\mathbf{x}$ located all around it. By selecting a kernel function K_a with bandwidth parameter a we have

$$\hat{\rho}(\mathbf{x}) = \frac{1}{na} \sum_{j=1}^{n} K_a(\mathbf{x}, \mathbf{x}^j).$$

The bandwidth a affects the roughness or smoothness of the estimate, and kernel based methods are sensitive to the choice of a.

Voronoi diagram based methods like [7] use the data all around $\mathbf{x}$ and the neighbourhood around $\mathbf{x}$ is thus obtained automatically.

2.2. K Nearest Neighbours and Natural Neighbours Estimators

The K nearest neighbours is a popular method in machine learning, see e.g. [15]. It is based on calculating the distances between the reference data (it is often called training data, although no actual training in the kNN method takes place) and the query point $\mathbf{x}$, at which either the value of a function or a class label is required.

Calculate the pairwise distances $d_i = \left\|\mathbf{x} - \mathbf{x}^i\right\|$ (in some norm), and sort the data set in the order of increasing d_i. There are many works dedicated to the choice of such a norm, see, e.g. [15, 16], which is a very hard and context dependent problem. In this study we assume it is the Euclidean norm. Then approximate $f(\mathbf{x})$ by $y = \sum_{i=1}^{k} w_i f\left(\mathbf{x}^i\right)$, where the weights w_i are determined usually by some non-increasing function $w_i = h\left(d_i\right)$, see [16, 17]. It was also proposed [18] to use the Induced Ordered Weighted Averaging functions (Induced OWA) instead of the weighted mean to aggregate the values $f\left(\mathbf{x}^i\right)$ and to learn $\mathbf{w}$ from the data. The Choquet integral was used for the same purpose in [19].

Unlike in function approximation, in the case of density estimation the values $f\left(\mathbf{x}^i\right)$ are not given but need to be estimated from the data set itself. One measure of density applicable to the kNN approach is the reciprocal of the pairwise distances, which we present in Section 3.

Another popular method of multivariate approximation is the natural neighbour scheme by Sibson [12, 20, 21]. The idea of this method is to build an interpolant whose value at $\mathbf{x}$ would depend on a few data points close to $\mathbf{x}$ at the same time distributed *all around* $\mathbf{x}$, see Figure 1. It favorably contrasts with the nearest neighbour methods in which only the distance from $\mathbf{x}$ matters.

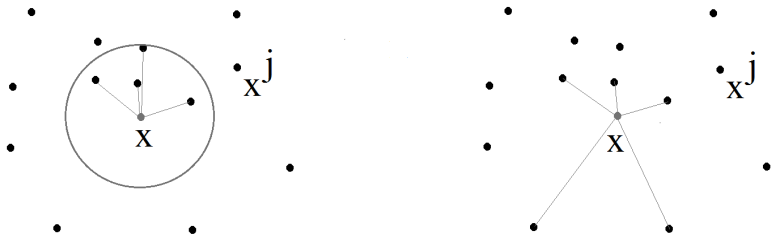

Figure 1. The nearest neighbours of a query point (left) versus natural neighbours (right)

In the natural neighbour scheme, the interpolant is a weighted average of the neighbouring data values

$$f(\mathbf{x}) = \sum_{j=1}^{J} w_j(\mathbf{x})\, f\left(\mathbf{x}^j\right),$$

where the weight $w_j(\mathbf{x})$ is proportional to the volume of the part of Voronoi cell $Vor(\mathbf{x}_j) = \{\mathbf{z} : \left\|\mathbf{z}-\mathbf{x}^j\right\| \leqslant \left\|\mathbf{z}-\mathbf{x}^k\right\|, k \neq j\}$, which is cut by the Voronoi cell $Vor(\mathbf{x}) = \left\{\mathbf{z} : \|\mathbf{z}-\mathbf{x}\| \leqslant \|\mathbf{z}-\mathbf{x}^k\|\right\}$, when $\mathbf{x}$ is added to the Voronoi diagram as one of the sites. Since Voronoi cell $Vor(\mathbf{x})$ borders only a few neighbouring Voronoi cells, only a few neighbouring data points *around* $\mathbf{x}$ participate in calculation of $f(\mathbf{x})$ (so called natural neighbours). More recently variations of Sibson's method were developed, based on other rules for calculating weights $w_j(\mathbf{x})$ [21, 22].

Sibson's interpolant possesses many useful properties, but it is computationally expensive, as each $\mathbf{x}$ requires computation of a new Voronoi diagram having $\mathbf{x}$ as one of the sites. There are methods that allow an update of the Voronoi diagram when $\mathbf{x}$ is added to the list of sites in 2- and 3-variate cases, so that the whole Voronoi diagram needs not be built for every $\mathbf{x}$. Such methods are very competitive, but we are unaware of any extension for more than three variables.

2.3. Fuzzy Measures and Discrete Choquet Integral

Aggregation of inputs into a representative output is the subject of aggregation functions [23, 24]. The weighted arithmetic mean (WAM) and the median are the two most commonly employed aggregation functions, and the WAM is used in the traditional kNN when averaging contributions of the K nearest neighbours. These functions are not suitable for our purpose as we want to account for input redundancies. The Choquet integral is a tool for explicitly modelling such interactions.

While the weights of the inputs in the WAM are associated with relative importances of each input, a discrete fuzzy measure allows one to assign importances to all possible groups of inputs, and thus offers a much greater flexibility for modeling aggregation.

Definition 1. Let $\mathcal{N} = \{1, 2, \ldots, n\}$. A discrete fuzzy measure is a set function $v : 2^{\mathcal{N}} \to [0, 1]$ which is monotonic (i.e. $v(\mathcal{A}) \leqslant v(\mathcal{B})$ whenever $\mathcal{A} \subset \mathcal{B}$) and satisfies $v(\emptyset) = 0$ and $v(\mathcal{N}) = 1$.

In Definition 1, a subset $\mathcal{A} \subseteq \mathcal{N}$ can be considered as a *coalition*, so that $v(\mathcal{A})$ gives us an idea about the importance or the weight of this coalition. The monotonicity condition implies that adding new elements to a coalition does not decrease its weight.

Definition 2. The discrete Choquet integral with respect to a fuzzy measure v is given by

$$C_v(\mathbf{x}) = \sum_{i=1}^{n} x_{(i)} \left[v\left(\left\{j | x_j \geqslant x_{(i)}\right\}\right) - v\left(\left\{j | x_j \geqslant x_{(i+1)}\right\}\right)\right], \tag{1}$$

where $\mathbf{x}_{\nearrow} = \left(x_{(1)}, x_{(2)}, \ldots, x_{(n)}\right)$ is a non-decreasing permutation of the input $\mathbf{x}$, and $x_{(n+1)} = \infty$ by convention.

Definition 3. Let v be a fuzzy measure. The Möbius transformation of v is a function defined for every $\mathcal{A} \subseteq \mathcal{N}$ as

$$\mathcal{M}(\mathcal{A}) = \sum_{\mathcal{B} \subseteq \mathcal{A}} (-1)^{|\mathcal{A} \setminus \mathcal{B}|} v(\mathcal{B}).$$

The WAM and ordered weighted averaging (OWA) functions are special cases of Choquet integrals with respect to additive and symmetric fuzzy measures respectively. In this contribution we are specifically interested in K-additive fuzzy measures [25, 26].

Definition 4. A fuzzy measure v is called K-additive ($1 \leqslant K \leqslant n$) if its Möbius transformation verifies

$$\mathcal{M}(\mathcal{A}) = 0$$

for any subset $\mathcal{A}$ with more than K elements, $|\mathcal{A}| > K$, and there exists a subset $\mathcal{B}$ with K elements such that $\mathcal{M}(\mathcal{B}) \neq 0$.

In this work we are interested in 2-additive fuzzy measures, therefore we assume all $\mathcal{M}(\mathcal{A}) = 0$ for $|\mathcal{A}| > 2$.

When dealing with multiple inputs, it is often the case that these are not independent, and there is some interaction (positive or negative) among the inputs. To measure such concepts as the importance of an input and interaction among the inputs we will use the concepts of Shapley value, which measures the importance of an input i in all possible coalitions, and the interaction index, which measures the interaction of a pair of inputs i, j in all possible coalitions [25, 26].

Definition 5. Let v be a fuzzy measure. The Shapley index for every $i \in \mathcal{N}$ is

$$\varphi(i) = \sum_{\mathcal{A} \subseteq \mathcal{N} \setminus \{i\}} \frac{(n - |\mathcal{A}| - 1)!|\mathcal{A}|!}{n!} \left[v(\mathcal{A} \cup \{i\}) - v(\mathcal{A})\right].$$

The Shapley value is the vector $\varphi(v) = (\varphi(1), \ldots, \varphi(n))$. It satisfies $\sum\limits_{i=1}^{n} \varphi(i) = 1$.

Definition 6. Let v be a fuzzy measure. The interaction index for every pair $i, j \in \mathcal{N}$ is

$$I_{ij} = \sum_{\mathcal{A} \subseteq \mathcal{N} \setminus \{i,j\}} \frac{(n - |\mathcal{A}| - 2)!|\mathcal{A}|!}{(n-1)!} \times \left[v(\mathcal{A} \cup \{i,j\}) - v(\mathcal{A} \cup \{i\}) - v(\mathcal{A} \cup \{j\}) + v(\mathcal{A})\right].$$

The interaction indices verify $I_{ij} < 0$ as soon as i, j are positively correlated (negative synergy). Similarly $I_{ij} > 0$ for negatively correlated inputs (positive synergy). $I_{ij} \in [-1, 1]$ for any pair i, j.

A fundamental property of K-additive fuzzy measures, which justifies their use in simplifying interactions between the criteria in multiple criteria decision making is the following [26].

Proposition 1. Let v be a K-additive fuzzy measure, $1 \leqslant K \leqslant n$. Then
– $I(\mathcal{A}) = 0$ for every $\mathcal{A} \subseteq \mathcal{N}$ such that $|\mathcal{A}| > K$;
– $I(\mathcal{A}) = \mathcal{M}(\mathcal{A})$ for every $\mathcal{A} \subseteq \mathcal{N}$ such that $|\mathcal{A}| = K$.

Thus K-additive measures acquire an interesting interpretation. These are fuzzy measures that limit interaction among the criteria to groups of size at most K. For instance, for 2-additive fuzzy measures, there are pairwise interactions among the criteria but no interactions in groups of 3 or more.

The Choquet integral can also be expressed in terms of interaction indices. For 2-additive fuzzy measures we have [27]:

$$C_I(\mathbf{x}) = \sum_{I_{ij}>0} \min(x_i, x_j)\, I_{ij} + \sum_{I_{ij}<0} \max(x_i, x_j)\, |I_{ij}| + \\ + \sum_{i=1\ldots K} x_i \left(\varphi(i) - \frac{1}{2} \sum_{i \neq j} |I_{ij}| \right), \qquad (2)$$

subject to

$$v(\{i\}) = \varphi(i) - \frac{1}{2} \sum_{i \neq j} |I_{ij}| \geqslant 0$$

for all $i = 1, \ldots, K$.

3. Nearest Neighbour Reweighted Graph

As we mentioned in the introduction, this density estimate is based on the kNN graph. Let us fix a value of K (sufficiently large to include the natural neighbours, of the order of tens to hundreds). Let us also fix a datum, $\mathbf{x}^j$ at which the density estimate will be computed. Calculate the pairwise distances from $\mathbf{x}^j$ to all the other pints in the sample and select the K nearest neighbours.

Let the density estimate at $\mathbf{x}^j$, $\hat{\rho}(\mathbf{x}^j)$ be given as a weighted sum of the values

$$\rho_{jk} = \frac{1}{||\mathbf{x}^j - \mathbf{x}^k||^d},$$

which are (up to a constant factor) the reciprocals of the volumes of spheres whose diameters are the segments between $\mathbf{x}^j$ and $\mathbf{x}^i$.

If we were to use a kNN estimate without reweighting, a large value of K would result in oversmoothing. Our goal is to select the weights in such a way that contributions of the neighbours on the same side relative to $\mathbf{x}^j$ are not double counted. This way only the natural neighbours all around $\mathbf{x}^j$ will contribute to the sum, and that would be equivalent to using a Delaunay based estimate but without its high complexity when d is large. The question is how to perform such weights redistribution.

Our main tool will be the discrete Choquet integral with respect to a fuzzy measure.

3.1. Construction of the Fuzzy Measure

We now construct such a fuzzy measure based on the proxies for interaction indices, which we call the redundancy values. In our setting the contributions of two neighbours, k and l, toward point density estimate are redundant if these neighbours lie on the same side from the query point $\mathbf{x}^j$. The degree of redundancy R_{kl} can be expressed

as a function of the cosine of the angle $\theta_{kl}^j = \angle \mathbf{x}^k \mathbf{x}^j \mathbf{x}^l$, which is easily computed as $\cos\left(\theta_{kl}^j\right) = \left(\mathbf{x}^j - \mathbf{x}^k\right) \cdot \left(\mathbf{x}^j - \mathbf{x}^l\right) / \left(\left\|\mathbf{x}^j - \mathbf{x}^k\right\| \left\|\mathbf{x}^j - \mathbf{x}^l\right\|\right)$. In other metric spaces R_{kl} can be computed without recurring to the scalar product, as a function of distances only.

Now, take the redundancy values $R_{kl} = g\left(\cos\left(\theta_{kl}^j\right)\right)$, where $g : [-1, 1] \to [0, 1]$ is some monotone function chosen as described below. Of course, the redundancy values cannot be taken as the (negative) interaction indices directly, because the interaction indices need to satisfy a number of constraints [25, p. 429], namely,

$$\frac{1}{2}\left(\sum_{j \in \mathcal{N} \setminus \mathcal{A} \cup \{i\}} I_{ij} - \sum_{l \in \mathcal{A}} I_{il}\right) \leqslant \varphi(i), \tag{3}$$

for all $\mathcal{A} \subseteq \mathcal{N} \setminus \{i\}$, $i = 1, \ldots, K$, where $\mathcal{N} = \{1, \ldots, K\}$ and $\varphi(i)$ are the Shapley indices. The constraints are satisfied if and only if v is a 2-additive fuzzy measure.

In addition, the Shapley values are also unknown. While it is possible to set up an optimization problem to select the interaction indices close to the redundancy values, but subject to the constraints (3), it would be extremely inefficient to solve such a problem for every datum $\mathbf{x}^j$. Instead we proceed as follows.

Let $C : [0, 1]^r \to [0, 1]$ be a triangular conorm [23], a monotone increasing symmetric associative function with neutral element 0. These functions are often used to aggregate inputs so that the total contribution does not exceed 1. The Einstein sum $C(x, y) = x + y - xy$ and the maximum function are prototypical examples of triangular conorms.

Let the initial contribution of all the K nearest neighbours of $\mathbf{x}^j$ be the same $w_k = 1/K$, $k = 1, \ldots, K$. Suppose that the neighbour $\mathbf{x}^l$ is located further than the inputs $k_1, k_2, \ldots, k_m$ and in roughly the same direction, so that $\theta_{k_1 l}^j, \ldots, \theta_{k_m l}^j$ are smaller than some threshold, like $\bar{\theta} = \pi/4$, see Figure 2. We want to redistribute the contribution from the input l to $k_1, \ldots, k_m$ proportionally to the redundancy values.

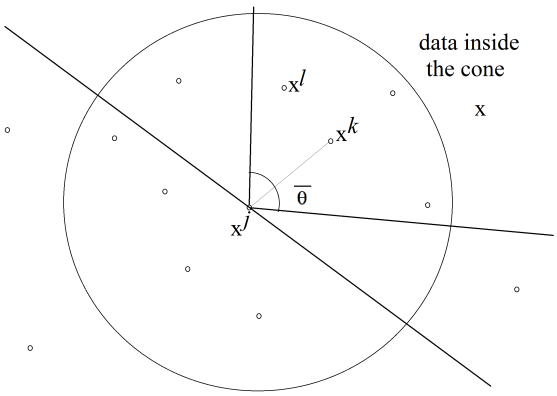

Figure 2. The contribution of inputs inside the cone is downweighted

We take the new weight $u_l = w_l \left(1 - C\left(R_{k_1 l}, \ldots, R_{k_m l}\right)\right)$. Note that $u_l \geqslant 0$ and becomes 0 only in case of at least one of the redundancy values $R_{k_i l} = 1$. The weights

of the inputs k_i are incremented by the value

$$w_{k_i} \to w_{k_i} + w_l \frac{R_{k_i l} C\left(R_{k_1 l}, \ldots, R_{k_m l}\right)}{\sum\limits_{t=1}^{m} R_{k_t l}}.$$

The weight w_l is updated $w_l \to u_l$.

By applying these formulas to every neighbour l from the furthest to the nearest, we downweight the contribution of the furthest and reallocate their weights to the nearer neighbours as long as those lie in the same direction (in the same spherical cone centered at $\mathbf{x}^j$ of angle of $\bar{\theta}$).

We can now state that the resulting reweighted sum $\sum\limits_{k=1}^{K} w_k \rho_k$ corresponds to the Choquet integral with respect to a 2-additive fuzzy measure whose interaction indices are negative and correspond to the redundancy values.

Theorem 1. *Let the redundancy values $0 \leqslant R_{kl} = R_{lk} \leqslant 1$, and let the weights be computed as*

$$w_k = \left(\frac{1}{K} + \sum_{t<k} w_t C_{m>t}\left(R_{mt}\right) \frac{R_{kt}}{\sum\limits_{m>t} R_{mt}} \right) \left(1 - C_{s>k}\left(R_{sk}\right)\right), \tag{4}$$

where $C_{\cdot}(\ldots)$ denotes the value of the triangular conorm applied to the arguments that satisfy the condition expressed in its subindex, analogously to the $\sum\limits_{\ldots}$ notation. Then the weighted sum $\sum\limits_{k=1}^{K} w_k \rho_k$ is equal to the Choquet integral of ρ_k with respect to some 2-additive fuzzy measure whose interaction indices I_{kl} are negative only when $R_{kl} > 0$.

Proof. Since the values of ρ_k are inversely proportional to the distances from $\mathbf{x}^j$ to $\mathbf{x}^k$, they are sorted in the order opposite to the order of $\mathbf{x}^k$. So we assume the neighbours are sorted in the order of decreasing distance to $\mathbf{x}^j$, and hence ρ_k are sorted in increasing order.

Consider the sequential process of calculating the weights w_k, $k = 1, \ldots, K$. Before the process starts all $w_k = 1/K$, which are positive and add to one. Take any iteration of this reweighting process, $k = q$, and assume that at its start all $w_k \geqslant 0$ and they add to one. We show that after that iteration is completed, the updated weights still add to one and are non-negative. We perform the following two steps

$$w_q \to w_q \left(1 - C_{k>q}\left(R_{kq}\right)\right),$$

and then for all $t > q$:

$$w_t \to w_t + w_q C_{k>q}\left(R_{kq}\right) \frac{R_{tq}}{\sum\limits_{k>q} R_{kq}}.$$

The value of $\sum\limits_{k} w_k$ does not change, as

$$w_q C_{k>q}\left(R_{kq}\right) - \sum_{t>q} w_q C_{k>q}\left(R_{kq}\right) \frac{R_{tq}}{\sum\limits_{k>q} R_{kq}} = 0,$$

and since the value of the triangular conorm C is no greater than one, w_q remains non-negative. Hence after the above iteration all w_k are still non-negative and add to one. By applying mathematical induction, these properties are maintained till the end of the iterative reweighting process. The formula (4) expresses the end result of the described reweighting process.

The weighted sum $\sum\limits_{k=1}^{K} w_k\rho_k$ can be expressed as the Choquet integral

$$\sum_{k=1}^{K} w_k\rho_k = \sum_{k=1}^{K} \rho_k \left(v\left(\{j|\rho_j \geqslant \rho_k\}\right) - v\left(\{j|\rho_j \geqslant \rho_{k+1}\}\right)\right) = C_v(\boldsymbol{\rho}) \tag{5}$$

with respect to some fuzzy measure v [23]. There are of course many such possible fuzzy measures, including additive and 2-additive measures, because we have only specified K out of 2^K fuzzy measure coefficients (in the form of w_k). In particular for the two-additive measure we have expression (2) [27].

In our case we discard the first sum as we only have to account for redundancies (all $I_{ij} \leqslant 0$) and hence our measure is submodular. We can therefore determine the values of $v(\{i\})$ and I_{ij} by matching the coefficients in (2), (5) with w_k, and setting $I_{ij} = 0$ whenever $R_{ij} = 0$. For this we obtain an underdetermined linear system of equations which always has at least one positive (in terms of the values $v(\{i\})$) solution. Furthermore we can set up a linear programming problem to maximize the values $v(\{i\})$ (in terms of their sum or their minimum) subject to matching (2), (5) with w_k, and the selected $I_{ij} \leqslant 0$ which always has a feasible solution (one of which is $v(\{i\}) = w_i$ and all $I_{ij} = 0$).□

So for the purposes of averaging local density values over the natural neighbours of $\mathbf{x}^j$ we fix K, triangular conorm C and a way of calculating the redundancy coefficients (from the cosines of the angles θ^j_{kl}), and then apply the iterative reweighting process expressed in (4) to calculate the density estimate $\hat{\rho}_j$ as the Choquet integral with respect to some submodular 2-additive fuzzy measure. In our experiments we used

$$R_{kl} = \max\left(0, \left(2\cos\left(\theta^j_{kl}\right)^2 - 1\right)\right)$$

for the threshold $\bar{\theta} = \pi/4$, and a modified version of this formula for other thresholds as described in the next section.

Three features of the reweighting method can be highlighted. Firstly, this method is equivariant to data translation, rotation and scaling (this property is expected from reliable estimators of density, mode and location). The reason is that the pairwise distances and angles used in calculations are not affected under these linear transformations. Secondly, the computational complexity of the presented algorithm is $O(dn^2 + dnK^2)$, based on the number of distance and angle calculations. Hence it will have performance gains over other natural neighbours schemata for larger dimensions, notably for $d \geqslant 8$. Thirdly, this method is fully parallelisable and also suitable for SIMD architectures like Graphics Processing Units (GPUs). Therefore quadratic complexity in n seems not to be much of an issue for $n \leqslant 10^6$.

3.2. Selection of the Threshold $\bar{\theta}$

We now discuss a method for choosing an appropriate value of the threshold $\bar{\theta}$ consistent across different dimensions d. If we choose $\bar{\theta} = \pi/2$, then the cone in which the data is assumed to be redundant becomes half-space in any dimension, so half of the nearest

neighbours of the point $\mathbf{x}^k$ are expected to be located in that half-space (assuming a locally uniform distribution). That may look too broad a choice, and one may select the redundant neighbours in a narrower cone, for example, choosing $\bar{\theta} = \pi/4$, see Figure 2. In the case of two-dimensional data such a cone will contain roughly a quarter of the nearest neighbours.

The difficulty is that when the dimension d increases, the probability that a near neighbour of $\mathbf{x}^k$ falls into such a cone of angle $\bar{\theta}$ decreases. This is due to the fact that in higher dimensions the volume of a spherical cone of angle $\bar{\theta} < \pi/2$ decreases compared to the volume of the ball. Therefore, in order for a spherical cone to contain approximately the same proportion of the near neighbours of a point across different dimensions we need to select the threshold $\bar{\theta}(d)$ as a function of the dimension of the space.

Let us consider the ratio of the volume of the intersection of a spherical cone with the ball of radius R to the volume of the ball $\mathrm{Vol}_c(d)/\mathrm{Vol}_s(d)$, the ratio we want to keep constant. With no loss of generality we can set $R = 1$.

It is known that

$$\mathrm{Vol}_s(d) = C_d R^d,$$

where the constant $C_d = \pi^{d/2}/\Gamma(1 + d/2)$ depends only on the dimension d. Γ is the standard gamma-function.

The spherical cone, which is the intersection of a cone C with the ball centered at the vertex of the cone can be represented as the union of two parts, the spherical cap (a non-empty intersection of a ball with a half-plane) and the intersection of the cone with the complement of the mentioned half-space, which we call the base cone B, see Figure 3.

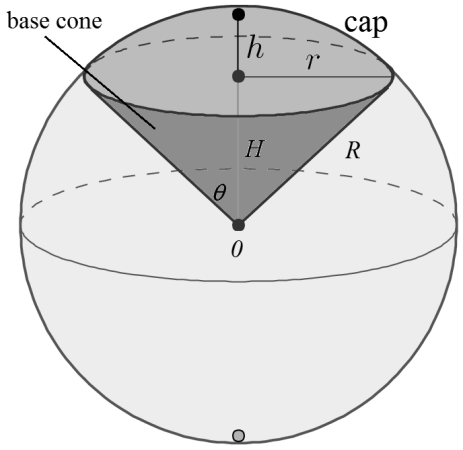

Figure 3. Three-dimensional spherical cone

It is also known that the volume of the spherical cap is given by [28]

$$\mathrm{Vol}_{\mathrm{cap}}(d) = \frac{1}{2} C_d R^d I_{(2Rh-h^2)/R^2}\left(\frac{d+1}{2}, \frac{1}{2}\right),$$

where $I_y(a, b)$ is the regularized incomplete beta function, and $h \leqslant R$ is the height of the cap.

Further, the volume of the base cone of height H and base radius r is given by

$$\mathrm{Vol}_{\mathrm{base}}(d) = \frac{Hr^{d-1}C_{d-1}}{d},$$

where $H = R - h$ and $r^2 = R^2 - H^2$. Therefore, assuming $R = 1$ and expressing $2h - h^2 = (1-H)(2-(1-H)) = \left(1-H^2\right)$, the volume of the spherical cone is

$$\mathrm{Vol}_c(d) = \frac{1}{2}C_d I_{(1-H^2)}\left(\frac{d+1}{2}, \frac{1}{2}\right) + \frac{H\left(1-H^2\right)^{\frac{d-1}{2}} C_{d-1}}{d}.$$

Now, let us fix the desired fraction of the volume of the ball $t = \mathrm{Vol}_c(d)/\mathrm{Vol}_s(d)$, for example $t = \frac{1}{4}$. Then we solve for H the equation

$$\frac{1}{2}I_{(1-H^2)}\left(\frac{d+1}{2}, \frac{1}{2}\right) + \frac{C_{d-1}}{C_d}\frac{H\left(1-H^2\right)^{\frac{d-1}{2}}}{d} = t.$$

From $H = \cos\left(\bar{\theta}(d)\right)$ we find the desired threshold $\bar{\theta}(d)$. The graph of $\cos\left(\bar{\theta}(d)\right)$ is presented on Figure 4. As expected, the first two values are $\cos\left(\bar{\theta}(2)\right) = 1/\sqrt{2}$ and $\cos\left(\bar{\theta}(3)\right) = 1/2$ which correspond to $\bar{\theta}(2) = \pi/4$ and $\bar{\theta}(3) = \pi/3$ respectively, but no closed form expression for the other values was found, although some simplifications using the relations between the gamma and beta functions can be made. Interestingly, the computed values are very well approximated by the function $g(d) = 1.2 - 0.839\tan^{-1}(\log(d))$, and this formula can be used for selecting a suitable value for the cosine of the threshold. The coefficients in the formula for g were obtained by the standard least squares regression with the approximation error RMSE= 0.004.

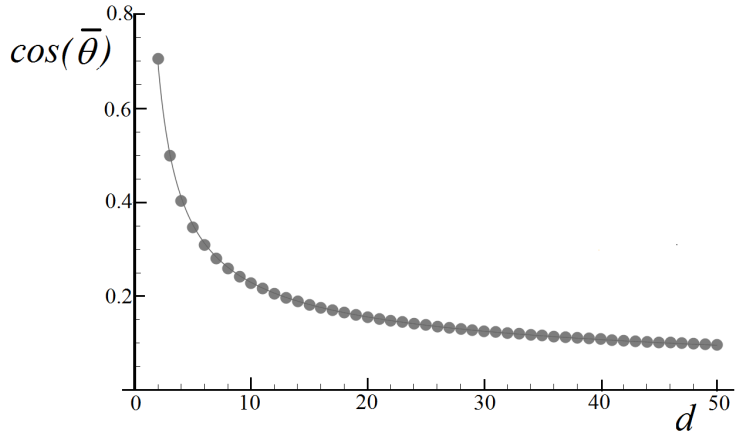

Figure 4. The graph of the values of the threshold as a function of the dimension $(d, \cos\left(\bar{\theta}(d)\right))$ and its approximating function g

4. Numerical Illustration

In order to show the advantages of the proposed method we will use highly structured data in $\mathbb{R}^d$ coming from a lower dimensional manifold. The reason is that if the data

are sampled from some standard test distribution, like a mixture of multivariate normals with nearly equal σ, the nearest neighbours of a point are distributed all around that point, and thus will overlap significantly with the set of natural neighbours we aim at identifying. In this case the proposed method shows quite similar results as the standard kNN and kernel estimates, provided that the value of K or the bandwidth are chosen appropriately to avoid oversmoothing.

It is for structured data that we expect significant benefits, i.e., when the nearest neighbours significantly differ from the natural neighbours. Furthermore, it turns out that this method is not sensitive to the choice of K, as contributions from the neighbours which are located beyond closer neighbours in the same direction are automatically downweighted.

Compared to Delaunay tessellation based methods, we expect to obtain computational advantages for higher dimensions. But for the purposes of illustration we limit ourselves to two-dimensional pictures. A detailed computational benchmarking is a subject for a followup paper.

Figure 5 presents a sample generated from a mixture (in equal proportions) of three products of normal distributions with parameters $(\mu_x, \mu_y, \sigma_x, \sigma_y)$ taken as $(0.13, 0.4, 0.08, 0.001)$, $(0.3, 0.3, 0.002, 0.8)$ and $(0.25, 0.4, 0.025, 0.025)$. Notice that the sample from the first distribution is practically located on a horizontal line, and because of the second component of the mixture spread in the other direction, standardization of the data does not alleviate this. The simulated mixture has three local modes at the centres of the above normal distributions, with the highest mode at $(0.13, 0.4)$ (notice small values of σ_x, σ_y for this component). The colour intensity of the data points in Figure 5 reflects the computed density at that point. The main mode of this mixture is at $(0.13, 0.4)$ and is significantly more pronounced than the two other modes. The sample size is 1000.

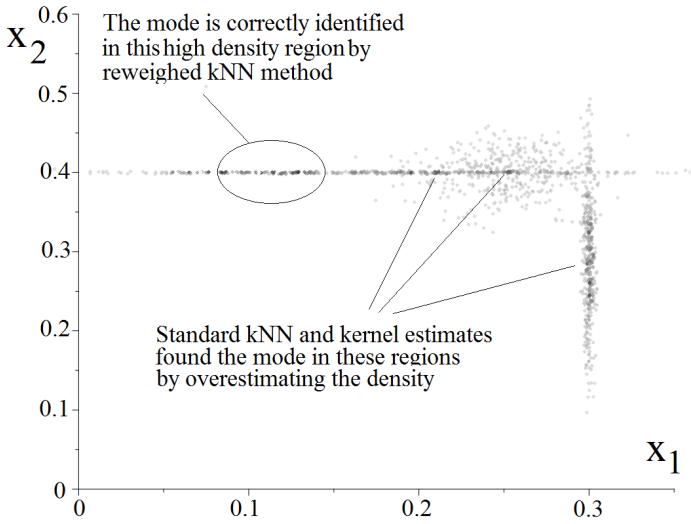

Figure 5. For this structured data sample the reweighted kNN shows advantages over other estimates which fail to identify the mode correctly

The reweighted kNN estimate (with $K = 150$, which is quite large) correctly identifies the regions of high density and points correctly to the mode. In contrast, the standard kNN with that value of K oversmoothens the estimate and incorrectly identifies the mode as that of the second component of the mixture. Other (smaller) values of K in the standard kNN incorrectly position the mode around $(0.25, 0.4)$. In fact, a careful manual adjustment to the value of K between 10 and 15 yields a better estimate of the mode at $(0.2, 0.4)$, but makes the estimate more "spiky" and overestimates the density at other places.

5. Conclusion

We have presented a multivariate density estimation method which calculates the local data density from the averaged distance to the natural neighbours of a point $\mathbf{x}$, the nearest neighbours distributed all around $\mathbf{x}$. The natural neighbours offer advantages over the standard kNN and kernel density estimates for structured data, for which the nearest neighbours could be distributed from one side of $\mathbf{x}$, thus introducing a bias into the estimate. However, the methods based on Voronoi and Delaunay tessellations, which compute the natural neighbours, suffer from high computational cost even for moderate dimension $d \geqslant 5$.

To alleviate prohibitive computational cost for higher dimensions we proposed a reweighting scheme, in which the contributions from a larger number of the nearest neighbours are reweighted based on their redundancy values, measured through the cosines of the angles these neighbours are visible from the point $\mathbf{x}$. These redundancy values serve as proxies for the interaction indices of a 2-additive fuzzy measure, with respect to which the pairwise distances are averaged by using the discrete Choquet integral. This way the contribution of the neighbours in the same direction as some of the nearer neighbours are discounted, and eventually only the contributions from the neighbours which all lie in distinct directions are accounted for. It is shown how redundancy values should be computed from the cosines of the angles in the multivariate setting according to the dimension d. The computational complexity of the proposed method is quadratic in the number of data n (same as the complexity of the kNN and kernel density estimates), and the method is fully parallelisable. Besides the kNN, the reweighting scheme can be used in conjunction with the kernel density estimates, which will be studied in the future.

We foresee applications of the proposed technique in density based clustering, mode estimation, image segmentation, anomaly detection and other areas of data analytics.

References

1. D. W. Scott, Multivariate Density Estimation, John Wiley and Sons, New York, 2015.
2. G. Beliakov, M. King, Density Based Fuzzy C-Means Clustering of Non-Convex Patterns, Europ. J. Oper. Res. 173 (2006) 717–728.
3. P. Angelov, R. R. Yager, Density-Based Averaging — a New Operator for Data Fusion, Information Sciences 222 (2013) 163–174.
4. G. Beliakov, T. Wilkin, On Some Properties of Weighted Averaging with Variable Weights, Information Sciences 281 (2014) 1–7.
5. E. Parzen, On the Estimation of a Probability Density Function and the Mode, Annals of Math. Stats. 33 (1962) 1065–1076.
6. C. Abraham, G. Biau, B. Cadre, Simple Estimation of the Mode of a Multivariate Density, The Canadian Journal of Statistics 31 (2003) 23–34.
7. W. E. Schaap, R. van de Weygaert, Continuous Fields and Discrete Samples: Reconstruction Through Delaunay Tessellations, Astronomy and Astrophysics 363 (2000) L29–L32.
8. E. Schubert, J. Sander, M. Ester, H. P. Kriegel, X. Xu, DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN, ACM Trans. Database Syst. 42 (2017) 19:1–19:21. doi:10.1145/3068335.
9. N.-B. Heidenreich, A. Schindler, S. Sperlich, Bandwidth Selection for Kernel Density Estimation: a Review of Fully Automatic Selectors, AStA Adv. Stat. 97 (2013) 403–433.
10. G. Voronoi, Nouvelles applications des parametres continus a la theorie des formes quadratiques, Journal fur die Reine und Angewandte Mathematik 133 (1908) 97–178.
11. B. Delaunay, Sur la sphère vide, Bulletin de l'Academie des Sciences de l'URSS, Classe des sciences mathematiques et naturelles 6 (1934) 793–800.
12. R. Sibson, Brief Description of Natural Neighbor Interpolation, in: V. Barnett (Ed.), Interpreting Multivariate Data, John Wiley and Sons, New York, 1981, pp. 21–36.

13. W. Stuetzle, Estimating the Cluster Tree of a Density by Analyzing the Minimal Spanning Tree of a Sample, Journal of Classification 20 (2003) 25–47.
14. H. Samet, Foundations of Multidimensional and Metric Data Structures, Elsevier, Boston, 2006.
15. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, New York, Berlin, Heidelberg, 2001.
16. B. Dasarathy, Nearest Neighbor Norms: NN Pattern Classification Techniques, IEEE Computer Society Press, Los Alamitos, CA, 1991.
17. S. Cost, S. Salzberg, A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features, Machine Learning 10 (1993) 57–78.
18. R. Yager, Using Fuzzy Methods to Model Nearest Neighbor Rules, IEEE Trans. on Syst., Man, and Cybernetics 32 (2002) 512–525.
19. E. Hüllermeier, The Choquet-Integral as an Aggregation Operator in Case-Based Learning, in: B. Reusch (Ed.), Computational Intelligence, Theory and Applications, Springer, Berlin, Heidelberg, 2006, pp. 615–627.
20. D. Watson, Contouring: A Guide to the Analysis and Display of Spatial Data, Pergamon Press, Oxford, 1992.
21. J.-D. Boissonnat, F. Cazals, Smooth Surface Reconstruction Via Natural Neighbour Interpolation of Distance Functions, Proc. of the 16th Annual Symposium on Computational Geometry (2000) 223–232.
22. V. V. Belikov, V. D. Ivanov, V. K. Kontorovich, S. A. Korytnik, A. Y. Semenov, The Non-Sibsonian Interpolation: a New Method of Interpolation of the Values of a Function on an Arbitrary Set of Points, Computational Mathematics and Mathematical Physics 37 (1997) 9–15.
23. G. Beliakov, A. Pradera, T. Calvo, Aggregation Functions: A Guide for Practitioners, Springer, Heidelberg, 2007.
24. M. Grabisch, J.-L. Marichal, R. Mesiar, E. Pap, Aggregation Functions, Cambridge University press, Cambridge, 2009.
25. M. Grabisch, T. Murofushi, M. Sugeno (Eds.), Fuzzy Measures and Integrals. Theory and Applications, Physica-Verlag, Heidelberg, 2000.
26. M. Grabisch, k-Order Additive Discrete Fuzzy Measures and Their Representation, Fuzzy Sets and Systems 92 (1997) 167–189.
27. B. Mayag, M. Grabisch, C. Labreuche, A Characterization of the 2-additive Choquet Integral, in: Proc. of IPMU, Malaga, Spain, 2008, pp. 1512–1518.
28. J. W. Harris, H. Stocker, Spherical Segment (Spherical Cap), in: Handbook of Mathematics and Computational Science, Springer, New York, 1998.

УДК 004.021, 004.8, 519.6
DOI: 10.22363/2312-9735-2018-26-1-58-73

Об одном методе оценки многомерной плотности на основе ближайших соседей

Глеб Беляков

Кафедра вычислительных технологий
Университет Дикин
Бурвуд хайвей 221, Бурвуд 3125, Австралия

Представлен метод оценки многомерной плотности, основанный на взвешенном методе ближайших соседей и имитирующий метод естественных соседей. Оценка многомерной плотности важна в машинном обучении, астрономии, биологии, физике и эконометрике. Строится 2-аддитивная нечёткая мера на основе аппроксимации индексов парных взаимодействий. Соседи, лежащие примерно в одном направлении, рассматриваются как излишние, и вклад дальнего соседа передаётся ближнему соседу. Расчёт локальной оценки плотности осуществляется с помощью дискретного интеграла Шоке таким образом, что учитывается

вклад соседей, расположенных со всех сторон точки, где производятся вычисления. Однако вклад соседей, расположенных с одной и той же стороны, занижается с помощью выбора подходящей нечёткой меры. Таким образом вычисляется приближение к множеству естественных соседей Сибсона. Этот метод значительно снижает вычислительную нагрузку методов на базе естественных соседей, которые лежат на основе тесселяции Делоне, в высокой размерности, для которых вычислительная сложность растёт как экспонента размерности. Описанный метод подходит для оценки плотности структурированных данных (возможно, лежащих на многообразии более низкой размерности), так как в этом случае ближайшие соседи могут значительно отличаться от естественных соседей.

Ключевые слова: оценка плотности, метод ближайших соседей, интеграл Шоке, нечёткая мера, метод естественных соседей

Литература

1. *Scott D. W.* Multivariate Density Estimation. — New York: John Wiley and Sons, 2015.
2. *Beliakov G., King M.* Density Based Fuzzy C-Means Clustering of Non-Convex Patterns // Europ. J. Oper. Res. — 2006. — Vol. 173. — Pp. 717–728.
3. *Angelov P., Yager R. R.* Density-Based Averaging — a New Operator for Data Fusion // Information Sciences. — 2013. — Vol. 222. — Pp. 163–174.
4. *Beliakov G., Wilkin T.* On Some Properties of Weighted Averaging with Variable Weights // Information Sciences. — 2014. — Vol. 281. — Pp. 1–7.
5. *Parzen E.* On the Estimation of a Probability Density Function and the Mode // Annals of Math. Stats. — 1962. — Vol. 33. — Pp. 1065–1076.
6. *Abraham C., Biau G., Cadre B.* Simple Estimation of the Mode of a Multivariate Density // The Canadian Journal of Statistics. — 2003. — Vol. 31. — Pp. 23–34.
7. *Schaap W. E., van de Weygaert R.* Continuous Fields and Discrete Samples: Reconstruction Through Delaunay Tessellations // Astronomy and Astrophysics. — 2000. — Vol. 363. — Pp. L29–L32.
8. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN / E. Schubert, J. Sander, M. Ester, H. P. Kriegel, X. Xu // ACM Trans. Database Syst. — 2017. — Vol. 42. — Pp. 19:1–19:21.
9. *Heidenreich N.-B., Schindler A., Sperlich S.* Bandwidth Selection for Kernel Density Estimation: a Review of Fully Automatic Selectors // AStA Adv. Stat. — 2013. — Vol. 97. — Pp. 403–433.
10. *Voronoi G.* Nouvelles applications des parametres continus a la theorie des formes quadratiques // Journal fur die Reine und Angewandte Mathematik. — 1908. — Vol. 133. — Pp. 97–178.
11. *Delaunay B.* Sur la sphère vide // Bulletin de l'Academie des Sciences de l'URSS, Classe des sciences mathematiques et naturelles. — 1934. — Vol. 6. — Pp. 793–800.
12. *Sibson R.* Brief Description of Natural Neighbor Interpolation // Interpreting Multivariate Data / Ed. by V. Barnett. — New York: John Wiley and Sons, 1981. — Pp. 21–36.
13. *Stuetzle W.* Estimating the Cluster Tree of a Density by Analyzing the Minimal Spanning Tree of a Sample // Journal of Classification. — 2003. — Vol. 20. — Pp. 25–47.
14. *Samet H.* Foundations of Multidimensional and Metric Data Structures. — Boston: Elsevier, 2006.
15. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. — New York, Berlin, Heidelberg: Springer-Verlag, 2001.
16. *Dasarathy B.* Nearest Neighbor Norms: NN Pattern Classification Techniques. — Los Alamitos, CA: IEEE Computer Society Press, 1991.
17. *Cost S., Salzberg S.* A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features // Machine Learning. — 1993. — Vol. 10. — Pp. 57–78.

18. *Yager R.* Using Fuzzy Methods to Model Nearest Neighbor Rules // IEEE Trans. on Syst., Man, and Cybernetics. — 2002. — Vol. 32. — Pp. 512–525.
19. *Hüllermeier E.* The Choquet-Integral as an Aggregation Operator in Case-Based Learning // Computational Intelligence, Theory and Applications / Ed. by B. Reusch. — Berlin, Heidelberg: Springer, 2006. — Pp. 615–627.
20. *Watson D.* Contouring: A Guide to the Analysis and Display of Spatial Data. — Oxford: Pergamon Press, 1992.
21. *Boissonnat J.-D., Cazals F.* Smooth Surface Reconstruction Via Natural Neighbour Interpolation of Distance Functions // Proc. of the 16th Annual Symposium on Computational Geometry. — 2000. — Pp. 223–232.
22. The Non-Sibsonian Interpolation: a New Method of Interpolation of the Values of a Function on an Arbitrary Set of Points / V. V. Belikov, V. D. Ivanov, V. K. Kontorovich, S. A. Korytnik, A. Y. Semenov // Computational Mathematics and Mathematical Physics. — 1997. — Vol. 37. — Pp. 9–15.
23. *Beliakov G., Pradera A., Calvo T.* Aggregation Functions: A Guide for Practitioners. — Heidelberg: Springer, 2007.
24. Aggregation Functions / M. Grabisch, J.-L. Marichal, R. Mesiar, E. Pap. — Cambridge: Cambridge University press, 2009.
25. Fuzzy Measures and Integrals. Theory and Applications / Ed. by M. Grabisch, T. Murofushi, M. Sugeno. — Heidelberg: Physica-Verlag, 2000.
26. *Grabisch M.* k-Order Additive Discrete Fuzzy Measures and Their Representation // Fuzzy Sets and Systems. — 1997. — Vol. 92. — Pp. 167–189.
27. *Mayag B., Grabisch M., Labreuche C.* A Characterization of the 2-additive Choquet Integral // Proc. of IPMU. — Malaga, Spain: 2008. — Pp. 1512–1518.
28. *Harris J. W., Stocker H.* Spherical Segment (Spherical Cap) // Handbook of Mathematics and Computational Science. — New York: Springer, 1998. — 107 p.

Для цитирования:
Beliakov Gleb On a Method of Multivariate Density Estimate Based on Nearest Neighbours Graphs // RUDN Journal of Mathematics, Information Sciences and Physics. — 2018. — Vol. 26, No 1. — Pp. 58–73. — DOI: 10.22363/2312-9735-2018-26-1-58-73.

For citation:
Beliakov Gleb On a Method of Multivariate Density Estimate Based on Nearest Neighbours Graphs, RUDN Journal of Mathematics, Information Sciences and Physics 26 (1) (2018) 58–73. DOI: 10.22363/2312-9735-2018-26-1-58-73.

Сведения об авторах:
Беляков Глеб — профессор, кандидат физико-математических наук, профессор кафедры вычислительных технологий Университета Дикин, Австралия (e-mail: `gleb@deakin.edu.au`, тел.: +61 3 925 17475)

Information about the authors:
Beliakov Gleb — professor, Candidate of Physical and Mathematical Sciences, professor of School of Information Technology of Deakin University, Australia (e-mail: `gleb@deakin.edu.au`, phone: +61 3 925 17475)