УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА.
СЕРИЯ ФИЗИКО-МАТЕМАТИЧЕСКИЕ НАУКИ

2018, Т. 160, кн. 2
С. 327–338

ISSN 2541-7746 (Print)
ISSN 2500-2198 (Online)

UDK 519.23

MANIFOLD LEARNING BASED ON KERNEL DENSITY ESTIMATION

A.P. Kuleshov[a], *A.V. Bernstein*[a,b], *Yu.A. Yanovich*[a,b,c]

[a]*Skolkovo Institute of Science and Technology, Moscow, 143026 Russia*
[b]*Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127051 Russia*
[c]*National Research University Higher School of Economics, Moscow, 101000 Russia*

Abstract

The problem of unknown high-dimensional density estimation has been considered. It has been suggested that the support of its measure is a low-dimensional data manifold. This problem arises in many data mining tasks. The paper proposes a new geometrically motivated solution to the problem in the framework of manifold learning, including estimation of an unknown support of the density.

Firstly, the problem of tangent bundle manifold learning has been solved, which resulted in the transformation of high-dimensional data into their low-dimensional features and estimation of the Riemann tensor on the data manifold. Following that, an unknown density of the constructed features has been estimated with the use of the appropriate kernel approach. Finally, using the estimated Riemann tensor, the final estimator of the initial density has been constructed.

Keywords: dimensionality reduction, manifold learning, manifold valued data, density estimation on manifold

Introduction

The general goal of data mining is to extract previously unknown information from the given dataset. Thus, it is supposed that the information is reflected in the structure of the dataset, which must be discovered by data analysis algorithms. Data mining faces a few main "super-problems", each associated with particular tasks: exploratory data analysis, clustering, classification, association pattern mining, outlier analysis, etc. These problems are challenging for data mining, because they act as building blocks in the context of a wide variety of data mining applications.

Smart mining algorithms are based on various data models, which reflect the dataset structure from algebraic, geometric, and probabilistic viewpoints and play the key role in data mining.

Geometrical models are motivated by the fact that many of the above tasks deal with real-world high-dimensional data. Furthermore, the "curse of dimensionality" phenomenon is often an obstacle to the use of many data analysis algorithms for solving these tasks.

Although data for the given data mining problem may have many features, in reality the intrinsic dimensionality of the data support (usually called data space, DS) of the full feature space may be low. It means that high-dimensional data form only a minor part in the high-dimensional "observation space", the intrinsic dimension of which is small. The most popular geometrical data model describing the low-dimensional structure of the DS is the manifold model [1], by which high-dimensional real-world data lie on

or near some unknown low-dimensional data manifold (DM) embedded in an ambient high-dimensional "observation" space. Various data analysis problems studied under this assumption about processed data, usually called manifold-valued data, are referred to as the manifold learning problems. Their general goal is to discover the low-dimensional structure of the high-dimensional DM from the given sample [2, 3].

Sampling models describe ways for extracting data from the DS. Typically, such models are a probabilistic: data are selected from the DS independently of each other according to an unknown probability measure on the DS, the support of which coincides with the DS. Statistical problems for the unknown probabilistic model consist in estimating an unknown probability measure or its various characteristics, including density. Notably, many high-dimensional data mining and analysis algorithms require accurate and efficient density estimators [4–11].

The paper considers a new geometrically motivated method for estimating an unknown density on the unknown low-dimensional DM based on the manifold learning framework and includes numerical experiments.

1. Density estimation on manifold: statement and related works

1.1. Assumptions about data manifold. Let $\mathbf{M}$ be an unknown "well-behaved" q-dimensional DM embedded in an ambient p-dimensional space R^p, $q \le p$; an intrinsic dimension q is assumed to be known. let us assume that the DM $\mathbf{M}$ is a compact manifold with the positive condition number [12]; thus, no self-intersections, no "short-circuit" are observed. For simplicity, we assume that the DM is covered by a single coordinate chart φ and, hence, has a form $\mathbf{M} = \{X = \varphi(b) \in R^p : b \in \mathbf{B} \subset R^q\}$, in which chart φ is one-to-one mapping from open bounded coordinate space $\mathbf{B} \subset R^q$ to the manifold $\mathbf{M} = \varphi(\mathbf{B})$ with inverse map $\psi = \varphi^{-1} : \mathbf{M} \to \mathbf{B}$. Inverse mapping ψ determines low-dimensional parameterization on the DM $\mathbf{M}$ (q-dimensional coordinates, or features, $\psi(X)$ of manifold points X), and chart φ recovers points $X = \varphi(b)$ from their features $b = \psi(X)$.

If the mappings $\psi(X)$ and $\psi(b)$ are differentiable (the covariant differentiation is used in $\psi(X)$, $X \in \mathbf{M}$) and $J_\psi(X)$ and $J_\varphi(b)$ are their $q \times p$ and $p \times q$ Jacobian matrices, respectively, then q-dimensional linear space

$$L(X) = \mathtt{Span}(J_\varphi(\psi(X))) \tag{1}$$

in R^p is a tangent space to the DM $\mathbf{M}$ at point $X \in \mathbf{M}$; hereinafter, $\mathtt{Span}(H)$ is a linear space spanned by the columns of arbitrary matrix H. These tangent spaces are considered as elements of the Grassmann manifold $\mathtt{Grass}(p, q)$ consisting of all q-dimensional linear subspaces in R^p.

As follows from identities $\varphi(\psi(X)) \equiv X$ and $\psi(\varphi(b)) \equiv b$ for all points $X \in \mathbf{M}$ and $b \in \mathbf{B}$, Jacobian matrices $\mathbf{J}_\psi(X)$ and $\mathbf{J}_\varphi(b)$ satisfy the relations $J_\varphi(\psi(X)) \times J_\psi(X) \equiv \pi(X)$ and $J_\psi(\varphi(b)) \times J_\varphi(b) \equiv I_q$, where I_q is a $q \times q$ unit matrix and $\pi(X)$ is $p \times p$ a projection matrix onto the tangent space $L(X)$ (1) to the DM $\mathbf{M}$ at point $X \in \mathbf{M}$.

Let us consider tangent space $L(X)$, in which point X corresponds to zero vector $\mathbf{0} \in L(X)$. Then, any point $Z \in L(X)$ can be expressed in polar coordinates as vector $t \times \theta$, where $t \in [0, \infty)$ and $\theta \in S_{q-1} \subset L(X)$, where S_{q-1} is the $(q-1)$-dimensional sphere in R^q.

Let us denote $\exp_X$, an exponential mapping from the $L(X)$ to the DM $\mathbf{M}$ defined in the small vicinity of the point $\mathbf{0} \in L(X)$. The inverse mapping $\exp_X^{-1}$ determines Riemann normal coordinates $t \times \theta = \exp_X^{-1}(X') \in R^q$ of near point $X' = \exp_X(t \times \theta)$.

1.2. Data manifold as Riemann manifold Let $Z = J_\varphi(\psi(X)) \times z$ and $Z' = J_\varphi(\psi(X)) \times z'$ be the vectors from tangent space $L(X)$ with coefficients $z \in R^q$ and

$z' \in R^q$ of expansion of these vectors in a basis consisting of columns of Jacobian matrix $J_\varphi(\psi(X))$. An inner product (Z, Z') induced by the inner product on R^p equals to $z^T \times \triangle_\varphi(X) \times z$, here $q \times q$ matrix $\triangle_\varphi(X) = (J_\varphi(\psi(X)))^T \times J_\varphi(\psi(X))$ is metric tensor on the DM $\mathbf{M}$. Thus, $\mathbf{M}$ is Riemann manifold $(\mathbf{M}, \triangle_\varphi)$ with Riemann tensor $\triangle_\varphi(X)$ in each manifold point $X \in \mathbf{M}$ smoothly varying from point to point [13, 14]. This tensor induces an infinitesimal volume element on each tangent space, and, thus, a Riemann measure on the manifold

$$m(dX) = \sqrt{|\det \triangle_\varphi(X)|} \times d_L(X), \tag{2}$$

where $d_L X$ is a Lebesgue measure on the DM $\mathbf{M}$ induced by exponential mapping $\exp_X$ from the Lebesgue measure on the $L(X)$. We denote $\theta_X(X')$, the volume density function on $\mathbf{M}$, as the square-root of the determinant of the metric $\triangle$ expressed in the Riemann normal coordinates of the point $\exp_X^{-1}(X')$. Strict mathematical definitions of these notations are in [15–17].

1.3. Probability measure on data manifold. Let $\sigma(\mathbf{M})$ be the Borel σ-algebra of $\mathbf{M}$ (the smallest σ-algebra containing all the open subsets of $\mathbf{M}$) and μ be a probability measure on the measurable space $(\mathbf{M}, \sigma(\mathbf{M}))$, the support of which coincides with the DM $\mathbf{M}$. Let us assume that μ is absolutely continuous with respect to the measure m (2), and

$$F(X) = \mu(dX)/m(dX) \tag{3}$$

is its density that separates from zero and infinity uniformly in the $\mathbf{M}$. This measure induces probabilistic measure ν (a distribution of random vector $b = \psi(X)$) on full-dimensional space $\mathbf{B} = \psi(\mathbf{M})$ with the standard Borel σ-algebra with density $f(b) = d\nu/db = |\det_\varphi(\varphi(b))|^{1/2} \times F(\varphi(b))$, with respect to the Lebesgue measure db in R^q. Hence,

$$F(X) = \left| \det_\varphi(X) \right|^{-1/2} \times f(\psi(X)). \tag{4}$$

1.4. Density on manifold estimation problem. Let dataset $\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\}$ consist of manifold points, which are randomly and independently of each other sampled from the DM $\mathbf{M}$ according to an unknown probability measure μ. We suppose that the DM $\mathbf{M}$ is "well-sampled"; this means that the sample size n is sufficiently large.

Given the dataset $\mathbf{X}_n$, the problem is to estimate the density $F(X)$ (3), as well as to estimate its support $\mathbf{M}$. Estimation of the DM $\mathbf{M}$ means construction of a q-dimensional manifold $\widehat{\mathbf{M}}$ embedded in an ambient Euclidean space R^p, which meets manifold proximity property $\widehat{\mathbf{M}} \approx \mathbf{M}$ meaning small Hausdorff distance $d_H(\widehat{\mathbf{M}}, \mathbf{M})$ between these manifolds. The desired estimator $\widehat{F}(X)$ defined on the constructed manifold $\widehat{\mathbf{M}}$ should provide proximity $\widehat{F}(X) \approx F(X)$ for all points $X \in \widehat{\mathbf{M}}$.

1.5. Manifold learning: related works. The goal of manifold learning (ML) is to find a description of the low-dimensional structure of an unknown q-dimensional DM $\mathbf{M}$ from random sample $\mathbf{X}_n$ [18]. The term "to find a description" is not formalized in general, and it has different meanings depending on the researcher's understanding.

In computational geometry, this term means "to approximate (to reconstruct) the manifold": to construct an area $\mathbf{M}^*$ in R^p that is "geometrically" close to the DM $\mathbf{M}$

in a suitable sense (using some proximity measure between subsets, such as the Hausdorff distance [18]), without finding a low-dimensional parameterization on the DM, which is usually required in the machine learning tasks.

The ML problem in machine learning/data mining is usually formulated as the manifold embedding problem: given dataset $\mathbf{X}_n$, to construct a low-dimensional parameterization of the DM $\mathbf{M}$, which produces an embedding mapping

$$h : X \in \mathbf{M} \subset R^p \to y = h(X) \in \mathbf{Y}_h = h(\mathbf{M}) \subset R^q \tag{5}$$

from the DM $\mathbf{M}$ to a feature space (FS) $\mathbf{Y}_h$ preserving the specific geometrical and topological properties of the DM, such as local data geometry, proximity relations, geodesic distances, angles, etc. Various manifold embedding methods, such as linear embedding, Laplacian eigenmaps, Hessian eigenmaps, ISOMAP, etc., are proposed; see, for example, [2, 3] and other surveys.

Manifold embedding is usually the first step in various machine learning/data mining tasks, in which reduced features $y = h(X)$ are used in the reduced learning procedures instead of initial p-dimensional vectors X. If the mapping h preserves only specific properties of high-dimensional data, then substantial data losses are possible when using a reduced vector $y = h(X)$ instead of the initial vector X. To prevent these losses, mapping h must preserve as much available information contained in the high-dimensional data as possible [18]; this means the possibility to recover high-dimensional points X from their low-dimensional representations $h(X)$ with small recovery error, which can describe a measure of preserving the information contained in high-dimensional data. Thus, it is necessary to find a recovery mapping

$$g : y \in \mathbf{Y}_h \to X = g(y) \in R^p \tag{6}$$

from the FS $\mathbf{Y}_h$ to the ambient space R^P which, together with the embedding mapping h (5), ensures proximity

$$r_{h,g}(X) \equiv g(h(X)) \approx X \quad \forall\, X \in \mathbf{M}, \tag{7}$$

in which $r_{h,g}(X)$ is the result of successive applying of embedding and recovery mappings to a vector $X \in \mathbf{M}$.

The reconstruction error $\delta_{h,g}(X) = |X - r_{h,g}(X)|$ is a measure of quality of the pair (h, g) at a point $X \in \mathbf{M}$. This pair determines a q-dimensional recovered data manifold (RDM) $\mathbf{M}_{h,g} = \{X = g(y) \in R^p : y \in \mathbf{Y}_h \subset R^q\}$ embedded in R^p and parameterized by a single chart g defined on the FS $\mathbf{Y}_h$. An inequality $d_H(\mathbf{M}_{h,g}, \mathbf{M}) \le \sup\limits_{X \in \mathbf{M}} |r_{h,g}(X) - X|$ implies manifold proximity

$$\mathbf{M} \approx \mathbf{M}_{h,g} \equiv r_{h,g}(\mathbf{M}). \tag{8}$$

There are some (though a limited number of) methods for recovery the DM $\mathbf{M}$ from the FS $\mathbf{Y}_h$. For the specific linear manifold, the recovery can be easily found using the principal component analysis (PCA) technique. For nonlinear manifolds, the sample-based auto-encoder neural networks [19, 20] determine both the embedding and recovery mappings. The general method, which constructs a recovery mapping in the same manner as the locally linear embedding algorithm [21] constructs an embedding mapping, has been introduced in [22]. Manifold recovery based on the estimated tangent spaces to the DM $\mathbf{M}$ are used in local tangent space alignment [23] and Grassman & Stiefel eigenmaps (GSE) [24] algorithms.

Due to further reasons, the manifold recovery problem can include the requirement to estimate Jacobian matrix J_g of mapping g (6) by certain $p \times q$ matrix $G_g(y)$ providing proximity $G_g(y) \approx J_g(y)\ \ \forall\, y \in \mathbf{Y}_h$.

This estimator G_g allows estimating the tangent spaces $L(X)$ to the DM $\mathbf{M}$ by q-dimensional linear spaces $L_{h,g}(X) = \mathtt{Span}(G_g(h(X)))$ in R^p, which approximates the tangent space to the RDM $\mathbf{M}_{h,g}$ at the point $r_{h,g} \in \mathbf{M}_{h,g}$ and provides tangent proximity

$$L(X) \approx L_{h,g}(X) \quad \forall\, X \in \mathbf{M} \tag{9}$$

between these tangent spaces in some selected metric on the Grassmann manifold $\mathtt{Grass}(p, q)$.

In manifold theory [13, 14], the set composed of manifold points equipped by tangent spaces at these points is called the tangent bundle of the manifold. Thus, the manifold recovery problem, which includes recovery of its tangent spaces as well, is referred to as the tangent bundle manifold learning problem: to construct the triple (h, g, G_g), which, additionally to manifold proximity (7), (8), provides tangent proximity (9) [25].

Matrix G_g determines $q \times q$ matrix $\triangle_{h,g}(X) = G_g^T(h(X)) \times G_g(h(X))$ consisting of inner products between the columns of the matrix $G\_g(h(X))$ and considered as the metric tensor on the RDM $\mathbf{M}_{h,g}$.

In real manifold learning/data mining tasks, intrinsic manifold dimension q is usually unknown too, but this integer parameter can be estimated with high accuracy from the given sample [26–30]: an error of dimension's estimator proposed in [30] has rate $O(\exp(-c \times n))$, in which constant $c > 0$ does not depend on sample size n. For this reason, the manifold dimension is usually assumed to be known (or already estimated).

1.6. Density estimation: related works. Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables taking values in R^d and having density function $p(x)$. Kernel density estimation is the most widely-used practical method for accurate nonparametric density estimation. Starting with the works of Rosenblatt [31] and Parzen [32], kernel density estimators have the form

$$\widehat{p}(x) = \frac{1}{na^d} \sum_{i=1}^{n} K_d\left(\frac{x - X_i}{a}\right), \tag{10}$$

Here, kernel function $K(t_1, t_2, \ldots, t_d)$ is a non-negative boundedness function that satisfies certain properties, the main of which is $\int\limits_{R^d} K_d\,(t_1, t_2, \cdots, t_d)\, dt_1 dt_2 \cdots dt_d = 1$, and "bandwidth" $a = a_n$ is chosen to approach to zero at a suitable rate as the number n of data points increases. Optimal bandwidth is $a_n = O(n^{-1/(d+4)})$ that yields the optimal rate of convergence of the mean squared error (MSE) of the estimator $\widehat{p}$:

$$\mathrm{MSE}(\widehat{p}) = \int\limits_{R^d} |\widehat{p}\,(x) - p\,(x)|^2\, p(x)\, dx = O(n^{-4/(d+4)}).$$

Therefore, it is not acceptable to use the kernel estimators (10) with MSE of the order $O(n^{-4/(p+4)})$ is not acceptable for high dimensional data.

Various generalizations of the estimator (10) were proposed. For example, adaptive kernel estimators were introduced in work [33], in which bandwidth $a = a_n(x)$ in (10) depends on x and is the distance between x and the k-nearest neighbor of x among $X_1, X_2, \ldots, X_n$, and $k = k_n$ is a sequence of non–random integers, such that $\lim\limits_{n\to\infty} k_n = \infty$.

Kernel estimators generally known as q-dimensional Riemann manifold embedded in the p-dimensional ambient Euclidean space were for the first time proposed by Pelletier

[16]. Let us denote $d_{\triangle}(X, X')$ as the Riemann distance (the length of the smallest geodesic curve) between near points X and X' defined by the known Riemann metric tensor $\triangle$. The proposed estimator

$$\widehat{p}(x) = \frac{1}{na^q} \sum_{i=1}^{n} \frac{1}{\theta_{X_i}(X)} K_1 \left(\frac{d(X, X_i)}{a} \right), \tag{11}$$

under the bandwidth $a_n = O(n^{-1/(q+4)})$, has the MSE of the order $O(n^{-4/(q+4)})$ [16, 34], which is acceptable for high-dimensional manifold valued data.

The paper [17] generalizes the estimators (11) to the estimators with adaptive kernel bandwidth $a_n(x)$ (similarly to the work [35] for the Euclidean space), depending on x.

The estimator (11) assumes that the DM $\mathbf{M}$ is known in advance and that we have access to certain geometric quantities related to this manifold such as intrinsic distances $d_{\triangle}(X, X')$ between its points and the volume density function $\theta_X(X')$. Thus, the estimator (11) cannot be used directly in cases where data live on an unknown Riemann manifold of R^p.

The paper [36] proposes a more straightforward method that directly estimates the density of data as being measured in the tangent space, without assuming any knowledge of the quantities about the intrinsic geometry of the manifold, such as its metric tensor, geodesic distances between its points, its volume form, etc. The proposed estimator

$$\widehat{p}(x) = \frac{1}{na^q} \sum_{i=1}^{n} K_1 \left(\frac{d_E(X, X_i)}{a} \right), \tag{12}$$

in which the Euclidean distance (in R^p) $d_E(X, X')$ between the near manifolds points X and X' is used. Under $a_n = O(n^{-1/(q+4)})$, this estimator has also optimal MSE order $O(n^{-4/(q+4)})$.

2. Density on manifold estimation: solution

2.1. Proposed approach. The proposed approach is introduced in [37] and it consists of three stages:

1) solving the tangent bundle manifold learning problem which results in the solution $(h, g, G_g \approx J_g)$;

2) estimating the density $f(y)$ of random feature $y = h(X)$ defined on the FS $\mathbf{Y}_h = h(\mathbf{M})$ from feature sample $\mathbf{Y}_n = \{y_i = h(X_i), i = 1, 2, \ldots, n\}$;

3) calculating the desired estimator $\widehat{F}(X)$ using $f(y)$ and $(h, g, G_g \approx J_g)$.

2.2. GSE solution to the tangent bundle manifold Learning. The solution for tangent bundle manifold learning is given by the GSE algorithm [38–40] and consists of several steps:

1) applying local principal component analysis (PCA) to approximate the tangent spaces. $\mathbf{M}$ at points $X \in \mathbf{M}$;

2) kernel on manifold definition construction;

3) tangent manifold learning;

4) embedding mapping construction;

5) kernel on feature space construction;

6) constructing the recovery mapping and its Jacobian.

2.3. Density on the manifold estimation. Based on the representation (4) and estimated embedding mapping $h(X)$ and Riemannian tensor $\triangle_{h,g}(X)$, the estimator

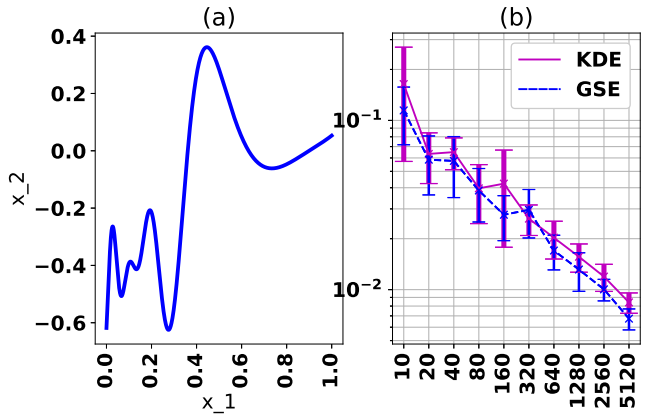

Fig. 1. (*a*) Manifold example; (*b*) MSE for $\widehat{p}$ (KDE, baseline method) and $\widehat{F}$ (GSE, proposed method)

$\widehat{F}(X)$ can be computed by the formula

$$\widehat{F}(X) = |\det \triangle_{h,g}(X)|^{1/2} \times \widehat{f}(h(X)). \tag{13}$$

The approximation $\triangle_{h,g}(X) \approx v^T(X) \times v(X)$, which yields equality

$$\left| \det_{h,g}(X) \right|^{1/2} \approx |\det(v(X))|\,,$$

allows us to simplify the estimator (13) to the following formula

$$\widehat{F}(X) = |\det(v(X))| \times \widehat{f}(h(X)).$$

3. Numerical experiments

The function $x_2 = \sin(30(x_1 - 0.9)^4)\cos(2(x_1 - 0.9)) + (x_1 - 0.9)/2,\ x_1 \in [0, 1]$, which was used in [41] to demonstrate a drawback of the kernel nonparametric regression (kriging) estimator with a stationary kernel (Fig. 1 (a)), was selected to compare the proposed kernel density estimator $\widehat{F}(X)$ (13) and stationary kernel density estimator $\widehat{p}(X)$ (12) in R^p. Here, $p = 2$, $q = 1$ and $X = (x_1, x_2)^T$. The kernel band-widths were optimized for both methods.

The same training data sets consisting of $n \in \{10, 20, 40, 80, 160, 320, 640, 1280, 2560, 5120\}$ points were used for constructing the estimators; the sample x_1 components were chosen randomly and uniformly distributed on the interval $[0, 1]$. The true probability was calculated theoretically. The errors were calculated for both estimators at the uniform grid on the interval with 100 001 points, then the mean squared errors (MSE) were calculated. The experiments were repeated $M = 10$ times, and the mean value of MSE and the mean plus/minus standard deviation are shown in Fig. 1, *b*. The numerical results show that the proposed approach is more suitable than the baseline algorithm.

Conclusions

The estimation problem for unknown density defined on the unknown manifold is solved within the manifold learning framework. A new geometrically motivated solution is proposed. The algorithm is a nonstationary kernel density estimator with a single

parameter for the kernel width. The numerical experiment with artificial data shows better results of the proposed approach against the ordinary kernel density estimator and could be considered as a proof of the concept example.

Acknowledgements. The study by A.V. Bernstein and Yu.A. Yanovich was supported by the Russian Science Foundation (project no. 14-50-00150).

References

1. Seung H.S. Cognition: The manifold ways of perception. *Science*, 2000, vol. 290, no. 5500, pp. 2268–2269. doi: 10.1126/science.290.5500.2268.
2. Huo X., Ni X.S., Smith A.K. A survey of manifold-based learning methods. In: *Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications*. Singapore, World Sci., 2008, pp. 691–745. doi: 10.1142/9789812779861_0015.
3. Ma Y., Fu Y. *Manifold Learning Theory and Applications*. London, CRC Press, 2011. 314 p.
4. Müller E., Assent I., Krieger R., Günnemann S., Seidl T. DensEst: Density estimation for data mining in high dimensional spaces. *Proc. 2009 SIAM Int. Conf. on Data Mining*. Philadelphia, Soc. Ind. Appl. Math., 2009. pp. 175–186, doi: 10.1137/1.9781611972795.16.
5. Kriegel H.P., Kroger P., Renz M., Wurst S. A generic framework for efficient subspace clustering of high-dimensional data. *Proc. 5th IEEE Int. Conf. on Data Mining (ICDM'05)*. Houston, TX, IEEE, 2005, pp. 250–257. doi: 10.1109/ICDM.2005.5.
6. Zhu F., Yan X., Han J., Yu P.S., Cheng H. Mining colossal frequent patterns by core pattern fusion. *Proc. 23rd IEEE Int. Conf. on Data Engineering*. Istanbul, IEEE, 2007, pp. 706–715. doi: 10.1109/ICDE.2007.367916.
7. Bradley P., Fayyad U., Reina C. Scaling clustering algorithms to large databases. *KDD-98 Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining*. New York, Am. Assoc. Artif. Intell., 1998, pp. 9–15.
8. Weber R., Schek H.J., Blott S. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. *Proc. 24th VLDB Conf.*. New York, 1998, pp. 194–205.
9. Domeniconi C., Gunopulos D. An efficient density-based approach for data mining tasks. *Knowl. Inf. Syst.*, 2004, vol. 6, no. 6, pp. 750–770. doi: 10.1007/s10115-003-0131-8.
10. Bennett K.P., Fayyad U., Geiger D. Density-based indexing for approximate nearest-neighbor queries. *KDD-99 Proc. 5th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. New York, 1999, pp. 233–243. doi: 10.1145/312129.312236.
11. Scott D.W. Multivariate density estimation and visualization. In: Gentle J.E., Härdle W.K., Mori Yu. (Eds.) *Handbook of Computational Statistics*. Berlin, Heidelberg, Springer, 2012, pp. 549–569. doi: 10.1007/978-3-642-21551-3.
12. Niyogi P., Smale S., Weinberger S. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 2008, vol. 39, nos. 1–3, pp. 419–441. doi: 10.1007/s00454-008-9053-2.
13. Jost J. *Riemannian Geometry and Geometric Analysis*. Berlin, Heidelberg, Springer, 2005. xiii, 566 p. doi: 10.1007/3-540-28891-0.
14. Lee J. *Manifolds and Differential Geometry*. Vol. 107: Graduate Studies in Mathematics. Am. Math. Soc., 2009. 671 p.
15. Pennec X. Probabilities and statistics on Riemannian manifolds: Basic tools for geometric measurements. *Int. Workshop on Nonlenear Signal and Image Processing (NSIP-99)*. Antalya, 1999, pp. 194–198.

16. Pelletier B. Kernel density estimation on Riemannian manifolds. *Stat. Probab. Lett.*, 2005, vol. 73, no. 3, pp. 297–304. doi: 10.1016/j.spl.2005.04.004.

17. Guillermo H., Munoz A., Rodriguez D. Locally adaptive density estimation on Riemannian manifolds. *Sort: Stat. Oper. Res. Trans.*, 2013, vol. 37, no. 2, pp. 111–130.

18. Freedman D. Efficient simplicial reconstructions of manifolds from their samples. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, vol. 24, no. 10, pp. 1349–1357. doi: 10.1109/TPAMI.2002.1039206.

19. Kramer M.A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.*, 1991, vol. 37, no. 2, pp. 233–243. doi: 10.1002/aic.690370209.

20. Dinh L., Sohl-Dickstein J., Bengio S. Density estimation using Real NVP. *arXiv:1605.08803*, 2016, pp. 1–32.

21. Zhang Z., Zha H. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Sci. Comput.*, 2004, vol. 26, no. 1, pp. 313–338. doi: 10.1137/S1064827502419154.

22. Bengio Y., Paiement J.-F., Vincent P. Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering. *Proc. 16th Int. Conf. on Neural Information Processing Systems*, 2003, pp. 177–184. doi: 10.1.1.5.1709.

23. Zhang P., Qiao H., Zhang B. An improved local tangent space alignment method for manifold learning. *Pattern Recognit. Lett.*, 2011, vol. 32, no. 2, pp. 181–189. doi: 10.1016/j.patrec.2010.10.005.

24. Bernstein A.V., Kuleshov A.P. Tangent bundle manifold learning via Grassmann & Stiefel eigenmaps. *arXiv:1212.6031*, 2012, pp. 1–25.

25. Bernstein A., Kuleshov A.P. Manifold Learning: Generalization ability and tangent proximity. *Int. J. Software Inf.*, 2013, vol. 7, no. 3, pp. 359–390.

26. Genovese C.R., Perone-Pacifico M., Verdinelli I., Wasserman L. Minimax manifold estimation. *J. Mach. Learn. Res.*, 2012, vol. 13, pp. 1263–1291.

27. Yanovich Yu. Asymptotic properties of local sampling on manifold. *J. Math. Stat.*, 2016, vol. 12, no. 3, pp. 157–175. doi: 10.3844/jmssp.2016.157.175.

28. Yanovich Yu. Asymptotic properties of nonparametric estimation on manifold. *Proc. 6th Workshop on Conformal and Probabilistic Prediction and Applications*, 2017, vol. 60, pp. 18–38.

29. Rozza A., Lombardi G., Rosa M., Casiraghi E., Campadelli P. IDEA: Intrinsic dimension estimation algorithm. *Proc. Int. Conf. "Image Analysis and Processing (ICIAP 2011)"*. Berlin, Heidelberg, Springer, 2011, pp. 433–442. doi: 10.1007/978-3-642-24085-0_45.

30. Campadelli P., Casiraghi E., Ceruti C., Rozza A. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Math. Probl. Eng.*, 2015, vol. 2015, art. 759567, pp. 1–21. doi: 10.1155/2015/759567.

31. Rosenblatt M. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.*, 1956, vol. 27, no. 3, pp. 832–837.

32. Parzen E. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 1962, vol. 33, no. 3, pp. 1065–1076.

33. Wagner T.J. Nonparametric estimates of probability densities. *IEEE Trans. Inf. Theory*, 1975, vol. 21, no. 4, pp. 438–440.

34. Henry G., Rodriguez D. Kernel density estimation on Riemannian manifolds: Asymptotic results. *J. Math. Imaging Vis.*, 2009, vol. 34, no. 3, pp. 235–239. doi: 10.1007/s10851-009-0145-2.

35. Hendriks H. Nonparametric estimation of a probability density on a Riemannian manifold using Fourier expansions. *Ann. Stat.*, 1990, vol. 18, no. 2, pp. 832–849.

36. Ozakin A., Gray A. Submanifold density estimation. *Proc. Conf. "Neural Information Processing Systems"(NIPS 2009)*, 2009, pp. 1–8.

37. Kuleshov A., Bernstein A., Yanovich Yu. High-dimensional density estimation for data mining tasks. *Proc. 2017 IEEE Int. Conf. on Data Mining (ICDMW)*. New Orleans, LA, IEEE, 2017, pp. 523–530. doi: 10.1109/ICDMW.2017.74.

38. Bernstein A., Kuleshov A., Yanovich Yu. Asymptotically optimal method for manifold estimation problem. *Proc. XXIX Eur. Meet. of Statisticians*. Budapest, 2013, pp. 8–9.

39. Kuleshov A., Bernstein A. Manifold learning in data mining tasks. *Proc. MLDM 2014: Machine Learning and Data Mining in Pattern Recognition*, 2014, pp. 119–133. doi: 10.1007/978-3-319-08979-9_10.

40. Bernstein A., Kuleshov A., Yanovich Yu. Information preserving and locally isometric & conformal embedding via Tangent Manifold Learning. *Proc. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Paris, IEEE, 2015, pp. 1–9. doi: 10.1109/DSAA.2015.7344815.

41. Xiong Y., Chen W., Apley D., Ding X. A non-stationary covariance-based Kriging method for metamodelling in engineering design. *Int. J. Numer. Methods Eng.*, 2007, vol. 71, no. 6, pp. 733–756. doi: 10.1002/nme.1969.

Received
October 17, 2017

Kuleshov Alexander Petrovich, Doctor of Technical Sciences, Professor, Academician of the Russian Academy of Sciences, Rector

Skolkovo Institute of Science and Technology
ul. Nobelya, 3, Territory of the Innovation Center "Skolkovo", Moscow, 143026 Russia
E-mail: *kuleshov@skoltech.ru*

Bernstein Alexander Vladimirovich, Doctor of Physical and Mathematical Sciences, Professor of the Center for Computational and Data-Intensive Science and Engineering; Leading Researcher of the Intelligent Data Analysis and Predictive Modeling Laboratory

Skolkovo Institute of Science and Technology
ul. Nobelya, 3, Territory of the Innovation Center "Skolkovo", Moscow, 143026 Russia

Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences
Bolshoy Karetny pereulok 19, str. 1, Moscow, 127051 Russia
E-mail: *a.bernstein@skoltech.ru*

Yanovich Yury Alexandrovich, Candidate of Physical and Mathematical Sciences, Researcher of the Center for Computational and Data-Intensive Science and Engineering; Researcher of the Intelligent Data Analysis and Predictive Modeling Laboratory; Lecturer of the Faculty of Computer Science

Skolkovo Institute of Science and Technology
ul. Nobelya, 3, Territory of the Innovation Center "Skolkovo", Moscow, 143026 Russia

Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences
Bolshoy Karetny pereulok 19, str. 1, Moscow, 127051 Russia

National Research University "Higher School of Economics"
ul. Myasnitskaya, 20, Moscow, 101000 Russia
E-mail: *yury.yanovich@iitp.ru*

УДК 519.23

Оценка плотности основанная на моделировании многообразий

А.П. Кулешов[1], *А.В. Бернштейн*[1,2], *Ю.А. Янович*[1,2,3]

[1]*Сколковский институт науки и технологий, г. Москва, 143026, Россия*
[2]*Институт проблем передачи информации Харкевича РАН, г. Москва, 127051, Россия*
[3]*Национальный исследовательский университет «Высшая школа экономики», г. Москва, 101000, Россия*

Аннотация

Рассматривается задача оценивания неизвестной многомерной плотности. Предполагается, что носителем меры является низкоразмерное многообразие (многообразие данных). Подобная задача возникает во многих разделах анализа данных. В работе предложен новое геометрически мотивированное решение в рамках парадигмы моделирования многообразий, включающее оценивание неизвестного носителя плотности.

Решение разбивается на два шага. Сначала оценивается многообразие и его касательное расслоение, в результате чего многомерные данные получают низкоразмерные описания, и оценивается Риманов тензор на многообразии данных. После этого производится непараметрическое ядерное оценивание неизвестной плотности в искусственном низкоразмерном пространстве. В завершении из полученной на предыдущем шаге оценки при помощи Риманова тензора строится итоговая оценка исходной неизвестной плотности.

Ключевые слова: снижение размерности, моделирование многообразий, оценка плотности на многообразии

Поступила в редакцию
17.10.17

Кулешов Александр Петрович, доктор технических наук, профессор, академик РАН, ректор

Сколковский институт науки и технологий
ул. Нобеля, д. 3, Территория Инновационного Центра «Сколково», г. Москва, 143026, Россия

E-mail: *kuleshov@skoltech.ru*

Бернштейн Александр Владимирович, доктор физико-математических наук, профессор Центра по научным и инженерным вычислительным технологиям для задач с большими массивами данных; ведущий научный сотрудник лаборатории интеллектуального анализа данных и предсказательного моделирования

Сколковский институт науки и технологий
ул. Нобеля, д. 3, Территория Инновационного Центра «Сколково», г. Москва, 143026, Россия

Институт проблем передачи информации им. А.А. Харкевича РАН
Большой Каретный переулок, д. 19, стр. 1, г. Москва, 127051, Россия

E-mail: *a.bernstein@skoltech.ru*

Янович Юрий Александрович, кандидат физико-математических наук, научный сотрудник Центра по научным и инженерным вычислительным технологиям для задач с большими массивами данных; научный сотрудник лаборатории интеллектуального анализа данных и предсказательного моделирования; старший преподаватель факультета компьютерных наук

Сколковский институт науки и технологий
ул. Нобеля, д. 3, Территория Инновационного Центра «Сколково», г. Москва, 143026, Россия

Институт проблем передачи информации им. А.А. Харкевича РАН
Большой Каретный переулок, д. 19, стр. 1, г. Москва, 127051, Россия
Национальный исследовательский университет «Высшая школа экономики»
ул. Мясницкая, д. 20, г. Москва, 101000, Россия
E-mail: *yury.yanovich@iitp.ru*

⟨ ***For citation*:** Kuleshov A.P., Bernstein A.V., Yanovich Yu.A. Manifold learning based on kernel density estimation. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Fiziko-Matematicheskie Nauki*, 2018, vol. 160, no. 2, pp. 327–338. ⟩

⟨ ***Для цитирования*:** *Kuleshov A.P., Bernstein A.V., Yanovich Yu.A.* Manifold learning based on kernel density estimation // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. – 2018. – Т. 160, кн. 2. – С. 327–338. ⟩