УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА.
СЕРИЯ ФИЗИКО-МАТЕМАТИЧЕСКИЕ НАУКИ

2018, Т. 160, кн. 2
С. 229–242

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

UDK 519.23

MANIFOLD LEARNING IN STATISTICAL TASKS

A.V. Bernstein

Skolkovo Institute of Science and Technology, Moscow, 143026 Russia
Kharkevich Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, 127051 Russia

Abstract

Many tasks of data analysis deal with high-dimensional data, and curse of dimensionality is an obstacle to the use of many methods for their solving. In many applications, real-world data occupy only a very small part of high-dimensional observation space, the intrinsic dimension of which is essentially lower than the dimension of this space. A popular model for such data is a manifold model in accordance with which data lie on or near an unknown low-dimensional data manifold (DM) embedded in an ambient high-dimensional space. Data analysis tasks studied under this assumption are referred to as the manifold learning ones. Their general goal is to discover a low-dimensional structure of high-dimensional manifold valued data from the given dataset. If dataset points are sampled according to an unknown probability measure on the DM, we face statistical problems on manifold valued data. The paper gives a short review of statistical problems regarding high-dimensional manifold valued data and the methods for solving them.

Keywords: data analysis, mathematical statistics, manifold learning, manifold estimation, density on manifold estimation, regression on manifolds

Introduction

For many centuries, the data analysis has been used for processing the results of observations (measurements) over real objects or of performed large-scale or computational experiments. Mathematical statistics develops mathematical methods for data analysis tasks when there is a mathematical (probabilistic) model of the processed data.

In recent decades, advances in information, computer, and telecommunications technologies have given the possibility of storage, fast searching, and processing of massive amounts of data, as well as rapid transmission of data through the communication channels and remote access to them. This phenomenon gave rise to the 'BigData' paradigm, which focuses on new technological possibilities of processing large data volumes and diversity. These new capabilities allowed formulating and solving fundamentally new scientific and applied data analysis problems and also gave rise to the new university and academic discipline called data science [1]. In data science, many mathematical and statistical tools are required to find fundamental principles behind the data, but data science has a different approach than that of classical mathematics, which uses mathematical models to fit data and to extract information [2].

The BigData phenomenon means usually not only big amounts of data, but also their high dimensionality [3]. For example, in image analysis/machine vision tasks [4], gray-scale image described with a resolution of $N \times N$ pixels is represented as N^2-dimensional vectors with components specifying light intensities at image pixels, and N varies from tens to thousands. Similar huge dimensionalities arise also in many other

applied areas 'with intensive use of data' (speech recognition, text mining, web search, etc.).

When the dimensionality of data is large, many theoretical and applied data analysis algorithms perform poorly due to a statistical and computational 'curse of dimensionality' (e.g., a collinearity or 'near-collinearity' of high-dimensional data causes difficulties when doing regression), 'empty space phenomenon', and other reasons [5].

For example, the minimax error in regression problem, in which at least s times differentiable unknown function depending on p-dimensional input is estimated from n-independent observations, cannot achieve a convergence rate faster than $n^{-s/(2s+p)}$ [6, 7] when nonparametric estimators are used [8]. In the density estimation problem, standard estimators (e.g., multidimensional version [9] of Parzen–Rosenblatt kernel estimators [10, 11]) in the p-dimensional case have mean squared errors of the order $O(n^{-4/(p+4)})$ [9].

Fortunately, in many applications, especially in imaging and medical ones, 'real-world' high-dimensional data obtained from 'natural' sources occupy only a very small part of the 'observation' space; in other words, an intrinsic dimension of the 'data support' is essentially lower than a dimension of the ambient high-dimensional space. This phenomenon results in that original high-dimensional data can be transformed into their lower-dimensional representations (features), which then are used in reduced data analysis procedures. The problem of finding such representations is usually referred to as the dimensionality reduction (DR) problem: given an input dataset

$$\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\} \subset \mathbf{X} \tag{1}$$

sampled from an unknown data space (DS) $\mathbf{X} \subset \mathbb{R}^p$, find an '$n$-point' embedding mapping

$$h_{(n)} : \mathbf{X}_n \subset \mathbb{R}^p \to \mathbf{Y}_n = h_{(n)}(\mathbf{X}_n) = \{y_1, y_2, \ldots, y_n\} \subset \mathbb{R}^q \tag{2}$$

of the sample $\mathbf{X}_n$ to a q-dimensional dataset $\mathbf{Y}_n$ (feature sample), $q < p$, which 'faithfully represents' the sample $\mathbf{X}_n$. The dimensionality q is either assumed to be known or estimated from the same sample (1). The term 'faithfully represents' is not formalized in general, and it is different in various DR methods due to choosing some optimized cost function $L_{(n)}(\mathbf{Y}_n \mid \mathbf{X}_n)$, which defines an 'evaluation measure' for the DR and reflects the desired properties of the mapping $h_{(n)}$ (2). As is pointed out in [12], a general view on the DR can be based on the 'concept of cost functions'.

If the DS $\mathbf{X}$ is an unknown q-dimensional affine linear space L in $\mathbb{R}^p$, principal component analysis (PCA) [13] estimates L by q-dimensional linear space L_{PCA} that minimizes over L a residual $\sum\limits_{i=1}^{n} |X_i - \mathrm{Pr}_L(X_i)|^2$, where Pr_L is a linear projector onto L.

The L_{PCA} passes through the sample mean $\bar{X}$ and is spanned by q eigenvectors of sample covariance matrix $\sum\limits_{i=1}^{n} (X_i - \bar{X}) \times (X_i - \bar{X})^T$ corresponding to q-largest eigenvalues. Then, coordinates $y \in \mathbb{R}^q$ of projection $Pr_L(X)$ on q-dimensional subspace L_{PCA} are taken as a low-dimensional representation of vector X.

However, if the DS $\mathbf{X}$ is a nonlinear one, only various heuristic nonlinear techniques, such as multidimensional scaling [14], auto-encoder neural networks [15–18], kernel PCA [19], and others, were proposed in the last century for the DR solution. Note that these methods are not based on any mathematical model of processed data.

For the first time, a model for high-dimensional data, called the manifold model [20], which occupies a small part of the observation space $\mathbb{R}^p$, appeared only in 2000 and became the most popular model for such data. This model assumes that data lie

on or near an unknown manifold (data manifold, DM) $\mathbf{M}$ of the lower dimension $q < p$ embedded in an ambient high-dimensional input space $\mathbb{R}^p$. Typically, this manifold assumption is satisfied for real-world high-dimensional data obtained from 'natural' sources.

Various data analysis tasks studied under a manifold assumption on processed data, which are called manifold valued data, are usually referred to as the manifold learning problems [21–24], the general goal of which is to discover a low-dimensional structure of the high-dimensional DM from the given training dataset $\mathbf{X}_n$ (1) sampled from the DM. If the dataset points are selected from the DM $\mathbf{M}$ independently of each other according to some unknown probability measure μ, we face statistical problems on manifold valued data.

The paper gives a short review of statistical analysis tasks on manifold valued data.

1. Assumptions on processed data

1.1. Assumptions on data manifold Let $\mathbf{M}$ be an unknown 'well-behaved' q-dimensional data manifold embedded in an ambient p-dimensional space $\mathbb{R}^p$, $q \le p$; an intrinsic dimension q is assumed to be known. Let us assume that the DM is a Riemann compact manifold with positive condition number [25]; thus, neither self-intersections, nor 'short-circuit' are observed. For simplicity, we assume that the DM is covered by a single coordinate chart φ and, hence, has the form

$$\mathbf{M} = \{X = \varphi(b) \in \mathbb{R}^p : b \in \mathbf{B} \subset \mathbb{R}^q\} \tag{3}$$

where chart φ is one-to-one mapping from the open bounded coordinate space $\mathbf{B} \subset \mathbb{R}^q$ to the manifold $\mathbf{M} = \varphi(\mathbf{B})$ with inverse mapping $\psi = \varphi^{-1} : \mathbf{M} \to \mathbf{B}$. 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 $b = \psi(X)$.

Note that pair $(\varphi, \mathbf{B})$ in (3) is determined up to arbitrary one-to-one mapping χ from the space $\mathbb{R}^q$ into itself – another pair $(\varphi^*, \mathbf{B}^*)$, in which $\varphi^*(b^*) = \varphi(\chi^{-1}(b^*))$ and $\mathbf{B}^* = \chi(\mathbf{B})$, gives another representation $\mathbf{M} = \varphi^*(\mathbf{B}^*)$ of the manifold $\mathbf{M}$ (3) and another low-dimensional features $b^* = \psi^*(X) = \chi(\psi(X))$ of manifold points.

If the mappings $\psi(X)$ and $\varphi(b)$ 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, than the q-dimensional linear space

$$L(X) = \mathtt{Span}(J_\varphi(\psi(X))) \tag{4}$$

in $\mathbb{R}^p$ is a tangent space to the DM $\mathbf{M}$ at point $X \in \mathbf{M}$; hereinafter, $\mathtt{Span}(H)$ is linear space spanned by columns of the arbitrary matrix H, which, in turn, form a basis in $L(X)$. These tangent spaces are considered as elements of the Grassmann manifold $\mathrm{Grass}(p, q)$ consisting of all q-dimensional linear subspaces in $\mathbb{R}^p$ [26].

Let $Z = J_\varphi(\psi(X)) \times z$ and $Z' = J_\varphi(\psi(X)) \times z'$ be two vectors from tangent space $L(X)$ with coefficients $z \in \mathbb{R}^q$ and $z' \in \mathbb{R}^q$ of their expansions in the columns of Jacobian matrix $J_\varphi(\psi(X))$. An inner product $(Z, Z') = z^T \times \Delta_\varphi(X) \times z'$ induced by the inner product in ambient space $\mathbb{R}^p$ is determined by $q \times q$ matrix $\Delta(X) = (J_\varphi(\psi(X)))^T \times J_\varphi(\psi(X))$ called metric tensor on the DM $\mathbf{M}$ in manifold point $X \in \mathbf{M}$ smoothly varying from point to point [27, 28]. This tensor $\Delta(X)$ induces an infinitesimal volume element on tangent space $L(X)$, and, thus, a Riemann measure on the manifold

$$m(dX) = \sqrt{|\det \Delta(X)|} \times \mathrm{mes}(dX) \tag{5}$$

where $\mathrm{mes}(dX)$ is a Lebesgue measure on the DM $\mathbf{M}$ [29–31].

1.2. Assumptions on probability measure of data. Let $\sigma(\mathbf{M})$ be a 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 $m(dx)$ (5) on $\mathbf{M}$, and

$$f(X) = \frac{\mu(dX)}{m(dX)} \tag{6}$$

is its density that separates from zero and infinity uniformly in the $\mathbf{M}$.

2. Statistical analysis of manifold valued data

2.1. General. Let dataset (1) be a random sample from unknown probability measure μ, the support $\mathtt{Supp}(\mu)$ of which is an unknown data manifold $\mathbf{M}$ (3) with an unknown intrinsic dimensionality q embedded in the ambient space $\mathbb{R}^p$, $q < p$. In the statistical framework, the goal of manifold learning is to make statistical inferences about the DM from the sample $\mathbf{X}_n$. We present below some typical examples of such statistical problems:

- intrinsic dimension estimation;
- low-dimensional parameterization of the data manifold;
- estimation of the data manifold;
- estimation of tangent spaces to the data manifold;
- estimation of density on the data manifold;
- regression on the manifold, and others,

the solutions of which are described shortly below.

2.2. Preliminaries and notations. Let us introduce some general concepts and notations, which are used in most manifold learning methods. For referent point $X \in \mathbf{M}$, let us denote $X_k(X) \in \mathbf{X}_n$ as a k-th nearest neighbor of the point X (i.e., $|X_1(X) - X| \le |X_2(X) - X| \le \cdots \le |X_n(X) - X|$).

Let $K_{E,\varepsilon}(X, X') = \mathbb{I}\{X' \in U(X, \varepsilon)\}$ be the Euclidean kernel. Here, $\mathbb{I}$ is the indicator function and $U(X, \varepsilon) = \{X' \in \mathbf{X}_n : |X' - X| < \varepsilon\}$. If the sample size n is sufficiently large, then for small ε and for not very large values k, the points $X' \in U(X, \varepsilon)$ and nearest neighbors $\{X_k(X)\}$ lie near the q-dimensional tangent space $L(X)$ (4).

Let us introduce a weighted undirected sample graph $\Gamma(\mathbf{X}_n)$ consisting of sample points $\{X_i\}$ as nodes. For the given ε (or k), edges in $\Gamma(\mathbf{X}_n)$ connect the points X_i and X_j only when $|X_i - X_j| < \varepsilon$ (or when these points are among k-nearest neighbors relative to each other).

2.3. Estimating an intrinsic dimensionality of the data manifold. Roughly speaking, the intrinsic dimension (ID) of a subset set $\mathbf{X} \subset \mathbb{R}^p$ is the minimum number $q = ID(\mathbf{X})$ of parameters needed to generate the subset description so that the information loss is minimized [32]. There are various strong definitions of intrinsic dimension (topological ID, Hausdorff ID, Kolmogorov capacity ID, Correlation ID, and others [33]), which give the same values for 'non-exotic' subsets.

For example, the Hausdorff–Besicovitch ID is defined as follows. Let a set E_r consist of 'half-open' cubes $\{Q = [k \times r, k \times r + r)^p, k = 0, \pm 1, \pm 2, \dots\}$ with edge r, and denote $N(\mathbf{X}, r) = \#\{Q \in E_r : \mathbf{X} \cap Q \ne \varnothing\}$. Let there exist numbers $q = q_{HB}(\mathbf{X})$ and $V = V_{HB}(\mathbf{X})$ called the Hausdorff–Besicovitch ID and the volume of the subset $\mathbf{X}$, respectively, such that $N(\mathbf{X}, r)/(V \times r^{-q}) \to 1$ as $r \to 0$ meaning also that $\mathbf{X}$ is measurable with respect to Jordan.

The statistical problem is to estimate the $ID(\mathbf{X})$ from the dataset (1) randomly sampled from the DS $\mathbf{X}$. Note that $ID(\mathbf{X}_n) = 0$ for the above ID definitions.

Various procedures are proposed for this problem [34–39]. In some of them [34–36], the $ID(\mathbf{M})$ is estimated under manifold assumption $\mathbf{X} = DM\ \mathbf{M}$ (3). For example, the maximum likelihood estimator

$$\widehat{q}(\mathbf{M}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{k-1}\sum_{j=1}^{k-1}\ln\frac{|X_k(X_i) - X_i|}{|X_j(X_i) - X_i|}\right)^{-1} \tag{7}$$

is proposed with the use of k nearest neighbors $\{X_j(X_i)\}$ of the sample points [34].

2.4. Low-dimensional parameterization of the data manifold. Given an intrinsic dimension q and sample $\mathbf{X}_n$, the statistical problem is to construct an embedding mapping $h : \mathbf{M} \subset \mathbb{R}^p \to \mathbf{Y} = h(\mathbf{M}) \subset \mathbb{R}^q$ from the DM $\mathbf{M}$ to the feature space (FS) $\mathbf{Y}$ by preserving the local data geometry, proximity relations, geodesic distances, angles, etc., of the DM. Most of the solutions to this problem [21–24] start from constructing the 'n-point' mapping $\mathbf{X}_n \to \mathbf{Y}_n = \{y_1, y_2, \ldots, y_n\}$ (2) by minimizing the selected cost function $L_{(n)}(\mathbf{Y}_n \mid \mathbf{X}_n)$. Following that, values $h_n(X)$ for out-of-sample points $X \in \mathbf{M} \,/\, \mathbf{X}_n$ are computed using some interpolation procedures.

For example, ISOmetric MAPping (ISOMAP) [40] preserves the data manifold geometry by capturing geodesic distances $\{D_{ij}\}$ between all pairs $\{(X_i, X_j)\}$ of sample points. Firstly, the geodesic distances D_{ij} are estimated by the lengths of shortest paths $\{d_{ij}\}$ between the nodes X_i and X_j in the graph $\Gamma(\mathbf{X}_n)$, a good quality of these estimators is proven [41]. Then, the feature sample $\mathbf{Y}_n$ is constructed using multidimensional scaling [14] by minimizing the cost function

$$L_{MDS}(\mathbf{Y}_n \mid \mathbf{X}_n) = \sum_{i,j=1}^{n}\left(d_{ij}^2 - |y_i - y_j|^2\right)^2. \tag{8}$$

In Laplacian eigenmaps [42], feature sample $\mathbf{Y}_n$ minimizes the cost function

$$L_{LE}(\mathbf{Y}_n \mid \mathbf{X}_n) = \sum_{i,j=1}^{n} K_{E,\varepsilon}(X_i, X_j) \times \|y_i - y_j\|^2 \tag{9}$$

under the normalizing condition $\sum_{i,j=1}^{n} K_{E,\varepsilon}(X_i, X_j) \times (y_i \times y_i^T) = I_q$ required to avoid a degenerate solution; this approach preserves the intrinsic geometric structure of the DM.

In statistical framework, under an asymptotically small ε, the cost function (9) is a sampling analog of a quantity

$$F(h) = \int\limits_{\mathbf{M}} |\nabla_{\mathbf{M}} h(X)|^2\, \mu(dX) = \int\limits_{\mathbf{M}} (h \times \Delta_{\mathbf{M}} h)(X)\, \mu(dX) \tag{10}$$

called a Laplacian of the graph $\Gamma(\mathbf{X}_n)$, where $h(x)$ is some component of the continuous interpolated embedding mapping $y = h_n(X) \in \mathbb{R}^q$ defined on the DM $\mathbf{M}$, $\nabla_{\mathbf{M}} h$ is its covariant gradient and $\Delta_{\mathbf{M}} h$ is the Laplace–Beltrami operator on the DM. It was proven [43–45] that the components $h_{1,n}(X), h_{2,n}(X), \ldots, h_{q,n}(X)$ of the mapping $h_n(X)$ converge to the eigenfunctions $h_1(X), h_2(X), \ldots, h_q(X)$ of the Laplace-Beltrami operator $\Delta_{\mathbf{M}}$ corresponding to its smallest nonzero eigenvalues $\lambda_1 \le \cdots \le \lambda_q$.

Other examples of manifold parameterization algorithms are locally linear embedding [46], Hessian eigenmaps [47], Maximum variance unfolding [48], Manifold charting [49], etc.

2.5. Information preserving in low-dimensional parameterization. In applications, manifold parameterization is usually the first step in various data analysis tasks, where reduced q-dimensional features $y = h(X)$ are used in the reduced procedures instead of initial p-dimensional vectors X. If the embedding 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, the mapping h must preserve as much available information contained in the high-dimensional data as possible [50, 51].

This means the possibility to recover high-dimensional points X from their low-dimensional representations $y = h(X)$ using some recovering mapping $\mathbf{g}(y) : \mathbf{Y} \to \mathbb{R}^p$ with small recovery error $\delta_{h,g}(X) = |X - g(h(X))|$.

The mappings (h, g) determine the q-dimensional recovered data manifold (RDM)

$$\mathbf{M}_{h,g} = \{X = \mathbf{g}(y) \in \mathbb{R}^p : y \in \mathbf{Y} \subset \mathbb{R}^q\} \tag{11}$$

which is embedded in an ambient space $\mathbb{R}^p$ and covered by a single chart g. A small recovery error implies proximity $\mathbf{M}_{h,g} \approx \mathbf{M}$ between the manifolds meaning a small Hausdorff distance $d_H(M_{h,g}, M)$ due inequality $d_H(\mathbf{M}_{h,g}, \mathbf{M}) \le \sup\limits_{X \in \mathbf{M}} \delta_{h,g}(X)$.

Let $J_g(y)$ be a $p \times q$ Jacobian matrix of the recovery mapping g, then q-dimensional linear space $L_{h,g}(X) = \mathtt{Span}(J_g(h(X)))$ in $\mathbb{R}^p$ is a tangent space to the RDM $\mathbf{M}_{h,g}$ at the point $g(h(X)) \in \mathbf{M}_{h,g}$.

There are some (though a limited number of) methods for recovery DM $\mathbf{M}$ from the $\mathbf{Y} = h(\mathbf{M})$. For linear manifolds, the recovery can be easily found using the PCA [13]. For nonlinear manifolds, the sample-based auto-encoder neural networks [15–18] determine both the embedding and recovering mappings. The general method, which constructs a recovering mapping in the same manner as locally linear embedding [46] constructs an embedding mapping, has been introduced in [52]. An interpolation-like nonparametric regression reconstruction method for manifold recovery is used in the manifold learning procedure called local tangent space alignment [53]. The Grassman & Stiefel eigenmaps algorithm [54, 55] also solves the manifold recovery problem.

2.6. Estimation of the data manifold. The problem is to construct q-dimensional manifold $\mathbf{M}_n$ which estimates (or approximates, recovers) the DM $\mathbf{M}$ from the sample $\mathbf{X}_n$.

In the papers, related to computational geometry, this problem is formulated as follows: given the finite dataset $\mathbf{X}_n$, to construct some set $\mathbf{M}^* \subset \mathbb{R}^p$ that approximates $\mathbf{M}$ in a suitable sense [56]. The solutions to this problem are usually based on decomposition of the DM $\mathbf{M}$ in small regions (using, for example, Voronoi decomposition or Delaunay triangulation on $\mathbf{M}$) and each region is piecewise approximated by some geometrical structure such as simplicial complex [56], tangential Delaunay complex [57], finitely many affine subspaces called 'flats' [58], k-means and k-flats [59], etc. However, such methods have a common drawback: they do not find a low-dimensional parameterization on the data manifold; such parameterization is usually required in the data analysis tasks which deal with high-dimensional data.

In the statistical framework, when the DM $\mathbf{M}$ (3) covered by a single chart is estimated, the solution $\mathbf{M}_n$ should be also a q-dimensional manifold covered by a single chart and, therefore, have the form (11).

The recovery error $\delta_{h,g}(X)$ can be directly computed at sample points $X \in \mathbf{X}_n$; for the out-of-sample point, X it describes the generalization ability of the solution (h, g) at a specific point X. The local lower and upper bounds are obtained for the maximum reconstruction error in a small neighborhood of an arbitrary point $X \in \mathbf{M}$ [55]. These

bounds are defined in terms of the distance between the tangent spaces $L(X)$ and $L_{h,g}(X)$ to the DM $\mathbf{M}$ and RDM $\mathbf{M}_{h,g}$ at the points X and $g(h(X))$, respectively, in some selected metric on the Grassmann manifold $\mathrm{Grass}(p,q)$.

It follows from the bounds that the greater the distances between these tangent spaces are, the lower the local generalization ability of the solution (h,g) becomes. Thus, it is natural to require that the solution (h,g) ensures not only the manifold proximity $\mathbf{M}_{h,g} \approx \mathbf{M}$, but also the tangent proximity $L_{h,g}(X) \approx L(X)$ for all points $X \in \mathbf{M}$. In the manifold theory [27, 28], the set composed of manifold points equipped by tangent spaces at these points is called the tangent bundle of the manifold. Thus, the problem of manifold recovery, which includes recovery of its tangent spaces too, is referred to as the tangent bundle manifold learning problem. The Grassman & Stiefel eigenmaps algorithm [54, 55] gives the solution to this problem and, under the asymptotic $n \to \infty$ and appropriate choice of algorithm parameters (for example, when ball radius $\varepsilon = \varepsilon_n$ in the kernel $K_{E,\varepsilon}(X,X')$ has the order of $O(n^{-1/(q+2)})$), the following rates of convergence

$$|X - g(h(X))| = O(n^{-2/(q+2)}), \quad d_{P,2}(L(X), L_{h,g}(X)) = O(n^{-1/(q+2)}) \tag{12}$$

hold true with high probability uniformly in points $X \in \mathbf{M}$ [60]. Here, $d_{P,2}$ is the projection 2-norm metric on the Grassmann manifold [61, 62] called the min correlation metric in statistics [63]. The term 'an event occurs with high probability' means that its probability exceeds the value $(1 - C_\alpha/n^\alpha)$ for any n and $\alpha > 0$, and the constant C_α depends only on α. The first rate in (12) coincides with the asymptotically minimax lower bound for the Hausdorff distance between the DM and RDM, which was set out in [64].

2.7. Estimation of tangent spaces to the data manifold. The simplest estimator for the tangent space $L(X)$ to the DM $\mathbf{M}$ is the linear space $L_{PCA}(X)$, which is a result of applying the PCA [13] to the local dataset $U(X,\varepsilon)$ when the threshold ε is small enough. Asymptotic properties of this estimator were studied in [44, 65–69]. Nonasymptotic analysis of the $L_{PCA}(X)$ of tangent spaces was performed in [70].

Eigenvectors of the local sample covariance matrix

$$\Sigma(X) = \sum_{i=1}^{n} K_{E,\varepsilon}(X, X_i) \times (X_i - X) \times (X_i - X)^T \tag{13}$$

which correspond to q-largest eigenvalues, form basis in the $L_{PCA}(X)$. However, these bases are not agreed by with each other and can be very different, even in close points. In various papers [53–55], the 'aligned' bases $\{H_1(X), H_2(X), \ldots, H_q(X)\}$ are constructed in the $L_{PCA}(X)$. To provide locally isometric and conformal properties of manifold parameterization, orthogonal aligned bases were constructed in [71].

2.8. Estimation of density on the data manifold. The problem of estimating the unknown density $f(X)$ (6) on the DM $\mathbf{M}$ was studied in a few papers [30, 31, 72–78]. A new geometrically motivated nonstationary kernel density estimator for the unknown density based on the Grassman & Stiefel eigenmaps algorithm [54, 55] is proposed in [79].

2.9. Regression on the data manifold. Let $Z = \Phi(X)$ be an unknown smooth mapping from its domain of the definition $\mathbf{M}$ lying in the input space $\mathbb{R}^p$ to the m-dimensional output space R^m; the domain of definition $\mathbf{M}$ is also assumed to be unknown. Given the input-output sample $\{(X_i, Z_i = \Phi(X_i),\ i = 1, 2, \ldots, n\}$, a common

regression problem is to estimate an unknown mapping Φ. When the domain of definition $\mathbf{M}$ is q-dimensional input manifold, $q < p$, we are talking about a regression on the manifold estimation problem.

When the input manifold $\mathbf{M}$ is known, i.e., its low-dimensional parameterization ψ in (3) is known, regression on the manifolds problems can be reduced to the classical multivariate regression problem (multi-output one if $m > 1$) [80–82]. Under unknown input manifold, various particular solutions to this problem were obtained [83–96]. A common regression problem consisting in estimation of an unknown mapping Φ, its Jacobian matrix, and unknown input manifold $\mathbf{M}$ is studied in the paper [97].

Conclusions

The paper describes various statistical problems regarding high-dimensional manifold valued data, such as estimating the data manifold (including estimation of its intrinsic dimension and tangent spaces, construction of low-dimensional parameterization of the manifold), estimating the density on the data manifold, regression on manifold tasks, etc. A short review of the possible solutions to these tasks is given.

Acknowledgements. This work was supported by the Russian Science Foundation (project no. 14-50-00150).

References

1. Cleveland W.S. Data science: An action plan for expanding the technical areas of the field of statistics. *Int. Stat. Rev.*, 2001, vol. 69, no. 1, pp. 21–26. doi: 10.1111/j.1751-5823.2001.tb00477.x.
2. Chen Li M., Su Z., Jiang B. *Mathematical Problems in Data Science. Theoretical and Practical Methods*. Springer, 2015. xviii, 213 p. doi: 10.1007/978-3-319-25127-1.
3. Donoho D.L. High-dimensional data analysis: The curses and blessings of dimensionality. *Proc. AMS Conf. on Math Challenges of the 21st Century*, 2000, pp. 1–33.
4. Pinoli J.-Ch. *Mathematical Foundations of Image Processing and Analysis*. Vols. 1, 2. John Wiley & Sons, 2014. Vol. 1: 464 p. Vol. 2: 496 p.
5. Verleysen M. Learning high-dimensional data. In: Ablameyko S. et al. (Eds.) *Limitations and Future Trends in Neural Computation*. IOS Press, 2003. pp. 141–162.
6. Stone Ch.J. Optimal rates of convergence for nonparametric estimators. *Ann. Stat.*, 1980, vol. 8, no. 6, pp. 1348–1360.
7. Stone Ch.J. Optimal global rates of convergence for nonparametric regression. *Ann. Stat.*, 1982, vol. 10, no. 4, pp. 1040–1053.
8. Wasserman L. *All of Nonparametric Statistics*. New York, Springer, 2007. xii, 270 p. doi: 10.1007/0-387-30623-4.
9. Cacoullos T. Estimation of a multivariate density. *Ann. Inst. Stat. Math.*, 1966, vol. 18, no. 1, pp. 179–189.
10. Rosenblatt M. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.*, 1956, vol. 27, no. 3, pp. 832–837.
11. Parzen E. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 1962, vol. 33, no. 3, pp. 1065–1076.
12. Bunte K., Biehl M., Hammer B. Dimensionality reduction mappings. *Proc. IEEE Symp. on Computational Intelligence and Data Mining (CIDM)*. Paris, IEEE, 2011, pp. 349–356. doi: 10.1109/CIDM.2011.5949443.

13. Jolliffe I.T. *Principal Component Analysis*. New York, Springer, 2002. xxx, 488 p. doi: 10.1007/b98835.

14. Cox T.F., Cox M.A.A. *Multidimensional Scaling*. London, Chapman and Hall/CRC, 2000. 328 p.

15. Hecht-Nielsen R. Replicator neural networks for universal optimal source coding. *Science*, 1995, vol. 269, no. 5232, pp. 1860–1863. doi: 10.1126/science.269.5232.1860.

16. Hinton G.E., Salakhutdinov R.R. Reducing the dimensionality of data with neural networks. *Science*, 2006, vol. 313, no. 5786, pp. 504–507. doi: 10.1126/science.1127647.

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

18. DeMers D., Cottrell G.W. Non-linear dimensionality reduction. *Proc. Conf. Adv. Neural Inf. Process. Syst. 5*. San Francisco, Morgan Kaufmann Publ., 1993, pp. 580–587.

19. Schölkopf B., Smola A., Müller K.-R., Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 1998, vol. 10, no. 5, pp. 1299–1319. doi: 10.1162/089976698300017467.

20. Seung H.S., Lee D.D. Cognition. The manifold ways of perception. *Science*, 2000, vol. 290, no. 5500, pp. 2268–2269. doi: 10.1126/science.290.5500.2268.

21. Cayton L. Algorithms for manifold learning. *Technical Report CS2008-0923*. Univ. of Calif. at San Diego, 2005, pp. 541–555, 2005.

22. Huo X., Ni X., Smith A.K. Survey of manifold-based learning methods. In: Liao T. Warren, Triantaphyllou E. *Recent Advances in Data Mining of Enterprise Data*. Singapore, World Sci., 2007, pp. 691–745. doi: 10.1142/6689.

23. Izenman A.J. Introduction to manifold learning. *Comput. Stat.*, 2012, vol. 4, no. 5, pp. 439–446. doi: 10.1002/wics.1222.

24. Ma Y., Fu Y. *Manifold Learning Theory and Applications*. London, CRC Press, 2011. 314 p.

25. 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.

26. Woods Yu.-Ch. Differential geometry of Grassmann manifolds. *Proc. Natl. Acad. Sci. USA*, 1967, vol. 57, no. 3, pp. 589–594.

27. Jost J. *Riemannian Geometry and Geometric Analysis*. Berlin, Heidelberg, Springer, 2011. xiii, 611 p. doi: 10.1007/978-3-642-21298-7.

28. Lee J.M. Manifolds and differential geometry. In: *Graduate Studies in Mathematics*. Vol. 107. Providence, R.I., Am. Math. Soc., 2009. 671 p.

29. Pennec X. Probabilities and statistics on Riemannian manifolds: Basic tools for geometric measurements. *J. Math. Imaging Vision*, 2006, vol. 25, no. 1, pp. 127–154. doi: 10.1007/s10851-006-6228-4.

30. 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.

31. Henry G., Muñoz A., Rodriguez D., Locally adaptive density estimation on Riemannian manifolds. *Stat. Oper. Res. Trans.*, 2013, vol. 37, no. 2, pp. 111–130.

32. Bennett R.S. The intrinsic dimensionality of signal collections. *IEEE Trans. Inf. Theory*, 1969, vol. 15, no. 5, pp. 517–525. doi: 0.1109/TIT.1969.1054365.

33. Katětov M., Simon P. Origins of dimension theory. In: Aull C.E., Lowen R. (Eds.) *Handbook of the History of General Topology*. Vol. 1. Dordrecht, Springer, 1997, pp. 113–134. doi: 10.1007/978-94-017-0468-7_11.

34. Levina E., Bickel P.J. Maximum likelihood estimation of intrinsic dimension. *Proc. 17th Int. Conf. on Neural Information Processing Systems*. Cambridge, MA, MIT Press, 2004, pp. 777–784.

35. Fan M., Qiao H., Zhang B. Intrinsic dimension estimation of manifolds by incising balls. *Pattern Recognit.*, 2009, vol. 42, no. 5, pp. 780–787. doi: 10.1016/j.patcog.2008.09.016.

36. Fan M., Gu N., Qiao H., Zhang B. Intrinsic dimension estimation of data by principal component analysis. *arXiv:1002.2050 [cs.CV]*, 2010, pp. 1–8.

37. Rozza A., Lombardi G., Rosa M., Casiraghi E., Campadelli P. IDEA: Intrinsic dimension estimation algorithm. *Proc. Int. Conf. on Image Analysis and Processing - ICIAP 2011. Lecture Notes in Computer Science*. Vol. 6978. Berlin, Heidelberg, Springer, 2011, pp. 433–442. doi: 10.1007/978-3-642-24085-0_45.

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

39. Camastra F., Staiano A. Intrinsic dimension estimation: Advances and open problems. *Inf. Sci.*, 2016, vol. 328, pp. 26–41. doi: 10.1016/j.ins.2015.08.029.

40. Tehenbaum J.B., Silva de V., Langford J.C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, vol. 290, no. 5500, pp. 2319–2323. doi: 10.1126/science.290.5500.2319.

41. Bernstein M., Silva de V., Langford J.C., Tenenbaum J.B. Graph approximations to geodesics on embedded manifolds. *Technical Report*. Stanford University, 2000. 26 p.

42. Belkin M., Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 2003, vol. 15, no. 6, pp. 1373–1396. doi: 10.1162/089976603321780317.

43. Rosasco L., Belkin M., De Vito E. On learning with integral operators. *J. Mach. Learn. Res.*, 2010, vol. 11, pp. 905–934.

44. Singer A., Wu. H-T. Vector diffusion maps and the connection Laplacian. *Commun. Pure Appl. Math.*, 2012, vol. 65, no. 8, pp. 1067–1144. doi: 10.1002/cpa.21395.

45. Yanovich Yu. Asymptotic properties of eigenvalues and eigenfunctions estimates of linear operators on manifolds. *Lobachevskii J. Math.*, 2017, vol. 38, no. 6, pp. 1–12.

46. Saul L.K., Roweis S.T. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, vol. 290, no. 5500, pp. 2323–2326. doi: 10.1126/science.290.5500.2323.

47. Donoho D.L., Grimes C. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA*, 2003, vol. 100, no. 10, pp. 5591–5596. doi: 10.1073/pnas.1031596100.

48. Weinberger K.Q., Saul L.K. Maximum variance unfolding: Unsupervised learning of image manifolds by semidefinite programming. *Int. J. Comput. Vision*, 2006, vol. 70, no. 1, pp. 77–90.

49. Brand M. Charting a manifold. *Proc. Int. Conf. on Advances in Neural Information Processing Systems*. Cambridge, MA, MIT Press, 2003, pp. 961–968.

50. Lee J.A., Verleysen M. Quality assessment of dimensionality reduction based on k-ary neighborhoods. *JMLR: JMLR Workshop and Conf. Proc.* Vol. 4: New challenges for feature selection in data mining and knowledge discovery. Antwerpen, 2008, pp. 21–35.

51. Lee J.A., Verleysen M. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 2009, vol. 72, nos. 7–9, pp. 1431–1443. doi: 10.1016/j.neucom.2008.12.017.
52. Saul L.K., Roweis S.T. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.*, 2003, vol. 4, no. 2, pp. 119–155. doi: 10.1162/153244304322972667.
53. Zhang Z., ZhaH. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM J. Sci. Comput.*, 2005, vol. 26, no. 1, pp. 313–338. doi: 10.1137/S1064827502419154.
54. Bernstein A.V., Kuleshov A.P. Tangent bundle manifold learning via Grassmann & Stiefel eigenmaps. *arXiv:1212.6031*, 2012, pp. 1–25.
55. Bernstein A.V., Kuleshov A.P. Manifold learning: Generalizing ability and tangent proximity. *Int. J. Software Inf.*, 2013, vol. 7, no. 3, pp. 359–390.
56. Freedman D. Efficient simplicial reconstructions of manifold from their samples. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, vol. 24, no. 10, pp. 1349–1357. doi: 10.1109/TPAMI.2002.1039206.
57. Boissonnat J.-D., Ghosh A. Manifold reconstruction using tangential delaunay complexes. *Discr. Comput. Geom.*, 2014, vol. 51, no. 1, pp. 221–267. doi: 10.1007/s00454-013-9557-2.
58. Karygianni S., Frossard P. Tangent-based manifold approximation with locally linear models. *Signal Process.*, 2014, vol. 104, pp. 232–247. doi: 10.1016/j.sigpro.2014.03.047.
59. Canas G.D., Poggio T., Rosasco L. Learning manifolds with k-means and k-flats, *arXiv:1209.1121*, 2013, pp. 1–19.
60. Kuleshov A., Bernstein A., Yanovich Y. Asymptotically optimal method in manifold estimation. *Abstr. 29th Eur. Meet. of Statisticians.* Budapest, 2013, p. 325.
61. Wang L., Wang X., Feng J. Subspace distance analysis with application to adaptive Bayesian algorithm for face recognition. *Pattern Recognit.*, 2006, vol. 39, no. 3, pp. 456–464. doi: 10.1016/j.patcog.2005.08.015.
62. Hamm J., Lee D.D. Grassmann discriminant analysis: A unifying view on subspace-based learning. *Proc. 25th Int. Conf. on Machine Learning (ICML-08)*. Helsinki, 2008, pp. 376–383. doi: 10.1145/1390156.1390204.
63. Hotelling H. Relations between two sets of variables. *Biometrika*, 1936, vol. 28, nos. 3–4, pp. 321–377. doi: 10.1093/biomet/28.3-4.321.
64. Genovese C.R., Perone-Pacifico M., Verdinelli I., Wasserman L. Minimax manifold estimation. *J. Mach. Learn. Res.*, 2012, vol. 13, pp. 1263–1291.
65. Achlioptas D. Random matrices in data analysis. *Proc. 15th Eur. Conf. on Machine Learning. Lecture Notes in Computer Science.* Vol. 3202. Pisa, Springer, 2004, pp. 1–8. doi: 10.1007/978-3-540-30116-5_1.
66. Tyagi H., Vural E., Frossard P. Tangent space estimation for smooth embeddings of riemannian manifold. *arXiv:1208.1065*, 2013, pp. 1–35.
67. Coifman R.R., Lafon S., Lee A.B., Maggioni M., Warner F., Zucker S. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci.*, 2005, vol. 102, no. 21, pp. 7426–7431. doi: 10.1073/pnas.0500334102.
68. 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.
69. 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.

70. Kaslovsky D.N., Meyer F.G. Non-asymptotic analysis of tangent space perturbation. *J. IMA*, 2014, vol. 3, no. 2, pp. 134–187. doi: 10.1093/imaiai/iau004.

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

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

73. 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.

74. 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.

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

76. Park H.S. Asymptotic behavior of kernel density estimation from a geometry viewpoint. *Commun. Stat. – Theory Methods*, 2012, vol. 41, no. 19, pp. 3479–3496. doi: 10.1080/03610926.2011.585009.

77. Kim Y.T., Park H.S. Geometric structures arising from kernel density estimation on Riemannian manifolds. *J. Multivar. Anal.*, 2013, vol. 114, pp. 112–126. doi: 10.1016/j.jmva.2012.07.006.

78. Berry T., Sauer T. Density estimation on manifolds with boundary. *Comput. Stat. Data Anal.*, 2017, vol. 107, pp. 1–17. doi: 10.1016/j.csda.2016.09.011.

79. 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.

80. Pelletier B. Nonparametric regression estimation on closed Riemannian manifolds. *J. Nonparametric Stat.*, 2006, vol. 18, no. 1, pp. 57–67. doi: 10.1080/10485250500504828.

81. Loubes J.M., Pelletier B. A kernel-based classifier on a Riemannian manifold. *Stat. Decis.*, 2008, vol. 26, no. 1, pp. 35–51. doi: 10.1524/stnd.2008.0911.

82. Steinke F., Hein M., Scholkopf B. Nonparametric regression between general Riemannian manifolds. *SIAM J. Imaging Sci.*, 2010, vol. 3, no. 3, pp. 527–563. doi: 10.1137/080744189.

83. Banerjee M., Chakraborty R., Ofori E., Okun M.S., Vaillancourt D., Vemuri B.C. A nonlinear regression technique for manifold valued data with applications to medical image analysis. *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, IEEE, 2016, pp. 4424–4432. doi: 10.1109/CVPR.2016.479.

84. Hinkle J., Muralidharan P., Fletcher P.T. Polynomial regression on Riemannian manifolds. *arXiv:1201.2395*, 2012, pp. 1–14.

85. Liu G., Lin Z., Yu Y. Multi-output regression on the output manifold. *Pattern Recognit.*, 2009, vol. 42, no. 11, pp. 2737–2743. doi: 10.1016/j.patcog.2009.05.001.

86. Shi X., Styner M., Lieberman J., Ibrahim J.G., Lin W., Zhu H. Intrinsic regression models for manifold-valued data. *Med. Image Comput. Comput.-Assist. Interv.*, 2009, vol. 12, pt. 2, pp. 192–199.

87. Kim H.J., Adluru N., Collins M.D., Chung M.K., Bendlin B.B., Johnson S.C., Davidson R.J., Singh V. Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images. *Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition*. Columbus, OH, IEEE, 2014, pp. 2705–2712. doi: 10.1109/CVPR.2014.352.

88. Kim H.J., Adluru N., Bendlin B.B., Johnson S.C., Vemuri B.C., Singh V. Canonical correlation analysis on Riemannian manifolds and its applications. In: *Fleet D., Pajdla T., Schiele B., Tuytelaars T. (Eds.) Computer Vision - ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*. Vol. 8690. Springer, 2014, pp. 251–267. doi: 10.1007/978-3-319-10605-2_17.

89. Bickel P., Li B. Local polynomial regression on unknown manifolds. In: *IMS Lect. Notes – Monogr. Ser.*Vol. 54: Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. 2007, pp. 177–186.

90. Aswani A., Bickel P., Tomlin C. Regression on manifolds: Estimation of the exterior derivative. *Ann. Stat.*, 2011, vol. 39, no. 1, pp. 48–81. doi: 10.1214/10-AOS823.

91. Cheng M.Y., Wu H.T. Local linear regression on manifolds and its geometric interpretation. *J. Am. Stat. Assoc.*, 2013, vol. 108, no. 504, pp. 1421–1434. doi: 10.1080/01621459.2013.827984.

92. Yang Y., Dunson D.B. Bayesian manifold regression. *arXiv:1305.0617*, 2014, pp. 1–40.

93. Einbeck J., Evers L. Localized regression on principal manifolds. *Proc. 25th Int. Workshop on Statistical Modeling (IWSM 2010)*. Glasgow, 2010, pp. 1–6.

94. Fletcher P.T. Geodesic regression on Riemannian manifolds. *Proc. Int. Workshop on Mathematical Foundations of Computational Anatomy (MFCA)*, 2011, pp. 75–86.

95. Fletcher P.T. Geodesic regression and the theory of least squares on Riemannian manifolds. *Int. J. Comput. Vision*, 2013, vol. 105, no. 2, pp. 171–185. doi: 10.1007/s11263-012-0591-y.

96. Banerjee M., Chakraborty R., Ofori E., Vaillancourt D., Vemuri B.C. Nonlinear regression on Riemannian manifolds and its applications to neuro-image analysis. In: *Lecture Notes on Computer Science*. Vol. 9349: Medical image computing and computer-assisted intervention, pt. I. Springer, Heidelberg, 2015, pp. 719–727.

97. Kuleshov A., Bernstein A. Nonlinear multi-output regression on unknown input manifold. *Ann. Math. Artif. Intell.*, 2017, vol. 81, nos. 1–2, pp. 209–240. doi: 10.1007/s10472-017-9551-0.

Received
October 17, 2017

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*

УДК 519.23

Статистические задачи моделирования многообразий

А.В. Бернштейн

Сколковский институт науки и технологий, г. Москва, 143026, Россия
Институт проблем передачи информации Харкевича РАН, г. Москва, 127051, Россия

Аннотация

Многие задачи анализа данных имеют дело с высокоразмерными данными, и феномен проклятия размерности является препятствием для использования многих методов для их решения. Во многих приложениях многомерные данные занимают лишь очень малую часть высокоразмерного пространства наблюдений, имеющую существенно меньшую размерность по сравнению с размерностью этого пространства. Модель многообразия для таких данных, в соответствие которой данные лежат на (или вблизи) неизвестного низкоразмерного многообразия данных, вложенного в охватывающее высокоразмерное пространство, является популярной моделью для таких данных. Задачи анализа данных, изучаемые в рамках этой модели, принято называть задачами моделирования многообразий, общая цель которых состоит в выявлении низкоразмерной структуры в лежащих на многообразии данных по имеющейся конечной выборке. Если точки выборки извлечены из многообразия в соответствии с неизвестной вероятностной мерой на многообразии данных, мы имеем дело со статистическими задачами на многообразии данных. Статья содержит обзор таких статистических задач и методов их решения.

Ключевые слова: анализ данных, математическая статистика, моделирование многообразий, оценка плотности на многообразиях, регрессия на многообразиях

Поступила в редакцию
17.10.17

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*

⟨ ***For citation*:** Bernstein A.V. Manifold learning in statistical tasks. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Fiziko-Matematicheskie Nauki*, 2018, vol. 160, no. 2, pp. 229–242. ⟩

⟨ ***Для цитирования*:** *Bernstein A.V.* Manifold learning in statistical tasks // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. – 2018. – Т. 160, кн. 2. – С. 229–242. ⟩