Low-Dimensional Data Representation in Data Analysis

Alexander Bernstein[1,2] and Alexander Kuleshov[1,2]

[1] Kharkevich Institute for Information Transmission Problems RAS, Moscow, Russia
kuleshov@iitp.ru
[2] National Research University Higher School of Economics (HSE), Moscow, Russia
abernstein@hse.ru

Abstract. Many Data Analysis tasks deal with data which are presented in high-dimensional spaces, and the 'curse of dimensionality' phenomena is often an obstacle to the use of many methods, including Neural Network methods, for solving these tasks. To avoid these phenomena, various Representation learning algorithms are used, as a first key step in solutions of these tasks, to transform the original high-dimensional data into their lower-dimensional representations so that as much information as possible is preserved about the original data required for the considered task. The above Representation learning problems are formulated as various Dimensionality Reduction problems (Sample Embedding, Data Manifold embedding, Data Manifold reconstruction and newly proposed Tangent Bundle Manifold Learning) motivated by various Data Analysis tasks. A new geometrically motivated algorithm that solves all the considered Dimensionality Reduction problems is presented.

Keywords: Machine Learning, Representation Learning, Dimensionality Reduction, Manifold Learning, Tangent Learning, Tangent Bundle Manifold Learning, Kernel methods.

1 Introduction

The goal of Data Analysis, which is a part of Machine Learning, is to extract previously unknown information from a dataset. Thus, it is supposed that information is reflected in the structure of a dataset which must be discovered from the data. Many Data Analysis tasks, such as Pattern Recognition, Classification, Clustering, Prognosis, Function reconstruction, and others, which are challenging for machine learning algorithms, deal with real-world data that are presented in high-dimensional spaces, and the 'curse of dimensionality' phenomena is often an obstacle to the use of many methods for solving these tasks.

To avoid these phenomena, various Representation learning algorithms are used as a first key step in solutions of these tasks. Representation learning (Feature extraction) algorithms transform the original high-dimensional data into their lower-dimensional representations (or features) so that as much information as possible is preserved about the original data required for the considered Data Analysis task.

After that, the initial Data Analysis task may be reduced to the corresponding task for the constructed lower-dimensional representation of the original dataset.

N. El Gayar et al. (Eds.): ANNPR 2014, LNAI 8774, pp. 47–58, 2014.

Of course, construction of the low-dimensional data representation for subsequent using in specific Data Analysis task must depend on the considered task, and success of machine learning algorithms generally depends on the data representation [1].

Representation (Feature) learning problems that consist in extracting a low-dimensional structure from high-dimensional data can be formulated as various Dimensionality Reduction (DR) problems, whose different formalizations depend on Data Analysis tasks considered further.

This paper is about DR problems in Data Analysis tasks. We describe a few key Data Analysis tasks that lead to different formulations of the DR: Sample Embedding for Clustering, Data Space (Manifold) embedding for Classification, Manifold Learning for Forecasting, etc. We also present a new geometrically motivated algorithm that solves all the considered DR problems.

The rest of the paper is organized as follows. Sections 2-5 contain definitions of various DR problems motivated by their subsequent using in specific Data Analysis tasks. The proposed DR solution is described in Section 6.

2 Sample Embedding Problem

One of the key Data Analysis tasks related to unsupervised learning is Clustering, which consists in discovering groups and structures in data that contain 'similar' (in one sense or another) sample points. Constructing a low-dimensional representation of original high-dimensional data for subsequent solution of the Clustering problem may be formulated as a specific DR problem, which will be referred to as the **Sample Embedding** problem and is as follows: Given an input dataset

$$\mathbf{X}_n = \{X_1, X_2, \dots, X_n\} \subset \mathbf{X}$$

randomly sampled from an unknown Data Space (DS) $\mathbf{X}$ embedded in a p-dimensional Euclidean space R^p, find an 'n-point' Embedding mapping

$$h_{(n)}: \mathbf{X}_n \subset R^p \to \mathbf{Y}_n = h_{(n)}(\mathbf{X}_n) = \{y_1, y_2, \dots, y_n\} \subset R^q \quad (1)$$

of the sample $\mathbf{X}_n$ to a q-dimensional dataset $\mathbf{Y}_n$ (feature sample), $q < p$, which 'faithfully represents' the sample $\mathbf{X}_n$ while inheriting certain subject-driven data properties like preserving the local data geometry, proximity relations, geodesic distances, angles, etc.

If the term 'faithfully represents' in the Sample Embedding problem corresponds to the 'similar' notion in the initial Clustering problem, we can solve the reduced Clustering problem for the constructed low-dimensional feature dataset $\mathbf{Y}_n$. After that, we can obtain some solution of the initial Clustering problem: clusters in the initial problem are images of clusters discovered in the reduced problem by using a natural inverse mapping from $\mathbf{Y}_n$ to the original dataset $\mathbf{X}_n$.

The term 'faithfully represents' is not formalized in general, and in various Sample Embedding methods it is different due to choosing some optimized cost function $L_{(n)}(\mathbf{Y}_n|\mathbf{X}_n)$ which defines an 'evaluation measure' for the DR and reflects desired

properties of the n-point Embedding mapping $h_{(n)}$ (1). As is pointed out in some papers, a general view on the DR can be based on the 'concept of cost functions.'

There exist a number of methods (techniques) for the Sample Embedding. Linear methods are well known and use such techniques as the PCA [2]. Various nonlinear techniques are based on Auto-Encoder Neural Networks [3, 4, 5], Kernel PCA [6], and others.

A newly emerging direction in the field of the Sample Embedding, which has been a subject of intensive research over the last decades, consists in constructing a family of algorithms based on studying the local structure of a given sampled dataset that retains local properties of the data with the use of various cost functions. Examples of such 'local' algorithms are: Locally Linear Embedding (LLE), Laplacian Eigenmaps (LE), Hessian Eigenmaps, ISOMAP, Local Tangent Space Alignment (LTSA), etc., described in [7, 8, 9] and other works. Some of these algorithms (LLE, LE, ISOMAP) can be considered in the same framework based on the Kernel PCA applied to various data-based kernels.

Note that Sample Embedding algorithms are based on the sample only, and no assumptions about the DS $\mathbf{X}$ are required for their descriptions. However, the study of properties of the algorithms is based on assumptions about both the DS and a way for extracting the sample from the DS.

3 Data Space (Manifold) Embedding problem

Another key Data Analysis task related to supervised learning concerns the Classification problem in which the original dataset consists of labeled examples: outputs (labels) $\mathbf{\Lambda}_n = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$ are known for the corresponding inputs $\{X_1, X_2, \dots, X_n\}$ sampled from the DS $\mathbf{X}$; each label λ belongs to a finite set $\{1, 2, \dots, m\}$ with $m \geq 2$. The problem is to generalize a function or mapping from inputs to outputs which can then be used to generate an output for a previously unseen input $X \in \mathbf{X}$.

In the case of high-dimensional original inputs $\mathbf{X}_n$, it is possible to construct low-dimensional features $\{y_1, y_2, \dots, y_n\}$ (1) by using the Sample Embedding algorithm. After that, we can consider the reduced sample $[\mathbf{Y}_n, \mathbf{\Lambda}_n]$ instead of the sample $[\mathbf{X}_n, \mathbf{\Lambda}_n]$. For the possibility of using the solution of the reduced classification problem built for the reduced dataset, it is necessary to construct a lower-dimensional representation for a new unseen (usually called Out-of-Sample, OoS) input $X \in \mathbf{X} / \mathbf{X}_n$. Thus, it is necessary to consider another specific DR problem which is an extension of the Sample Embedding and can be referred to as the Data Space Embedding (Parameterization) problem: Given an input dataset (sample) $\mathbf{X}_n$ from the DS $\mathbf{X} \subset R^p$, construct a low-dimensional parameterization of the DS which produces an Embedding mapping

$$h: \mathbf{X} \subset R^p \rightarrow \mathbf{Y} = h(\mathbf{X}) \subset R^q \tag{2}$$

from the DS $\mathbf{X}$, including the OoS points, to the Feature Space (FS) $\mathbf{Y} \subset R^q$, $q < p$, which preserves specific properties of the DS $\mathbf{X}$. The term 'preserves specific properties' is not

formalized in general and can be different due to choosing various cost functions reflecting specific preserved data properties.

The definition of the Data Space Embedding problem uses values of the Embedding mapping h (2) at the OoS points too. Thus, to justify the problem solution and study properties of the solution, we must define a Data Model describing the DS and a Sampling Model offering a way for extracting both the sample $\mathbf{X}_n$ and the OoS points from the DS. The most popular models in the DR are Manifold Data Models, see [7, 8, 9] and others works, in which the DS $\mathbf{X}$ is a q-dimensional manifold embedded in an ambient p-dimensional Euclidean space R^p, $q < p$, and referred to as the Data Manifold (DM). In most studies, DM is modeled using a single coordinate chart.

The Sampling Model is typically defined as a probability measure μ on the DM $\mathbf{X}$ whose support $\mathrm{Supp}(\mu)$ coincides with the DM $\mathbf{X}$. In accordance with this model, the dataset $\mathbf{X}_n$ and OoS points $X \in \mathbf{X} / \mathbf{X}_n$ are selected from the DM $\mathbf{X}$ independently of each other according to the probability measure μ.

A motivation for using the Manifold Data model consists in the following empirical fact: as a rule, high-dimensional real-world data lie on or near some unknown low-dimensional Data Manifold embedded in an ambient high-dimensional 'observation' space. This assumption is usually referred to as the Manifold assumption.

Various non-linear DR problems applied to the data which are described by the Manifold Data Model are usually referred to as the Manifold Learning (ML) problem [7, 8, 9]; the above-defined Data Space Embedding problem under the Manifold Data Model will be referred to as the Manifold Embedding problem. In the introduced terms, the Manifold Embedding problem is to construct a parameterization of the DM (global low-dimensional coordinates on the DM) from a finite dataset sampled from the DM. Note that there is no generally accepted definition for the ML.

Manifold assumption allowed constructing a family of algorithms based on studying the local structure of a given sampled dataset that retains local properties of the data with the use of various cost functions. Examples of such 'local' algorithms are described in [7, 8, 9] and other works; an 'OoS extension' for some local algorithms has been found in [10].

4 Manifold Learning Problem as Data Manifold Reconstruction

Manifold Embedding is usually a first step in various Data Analysis tasks in which reduced q-dimensional features $y = h(X)$ are used in the reduced learning procedures instead of initial p-dimensional vectors X. If the Embedding mapping h in the Manifold Embedding 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 as possible available information contained in the high-dimensional data [11]. Thus, it is necessary to consider the Manifold Embedding problem, in which the term *'faithfully represents'* has a specified meaning reflecting the possibility for reconstructing the initial vector $X \in \mathbf{X}$ from the feature $y = h(X)$ with small reconstruction error. Note that this error can be considered as a valid evaluation measure ('universal quality criterion') for Manifold Embedding procedures describing a measure of preserving information contained in the high-dimensional data [11].

There is a natural reconstruction of the vector $X \in \mathbf{X}$ from its lower-dimensional feature $y = h(X)$ for feature sample points $y \in \mathbf{Y}_n$. But in some tasks there may arise the problem of accurately reconstructing the points $X \in \mathbf{X}$ from their low-dimensional features $y = h(X)$ for Feature-Out-of-Sample, FOoS, points $y = h(X) \in h(\mathbf{X}) / \mathbf{Y}_n$. This possibility is directly required in various Data Analysis tasks such as multidimensional time series prognosis [12], data-based approximation of function with high-dimensional inputs [13], etc.

As an example, consider the problem of Electricity price curve forecasting [12] which is as follows. Electricity 'daily-prices' are described by a multidimensional time series (electricity price curve) $X_t = (X_{t1}, X_{t2}, \dots, X_{t,24})^T \in R^{24}$ consisting of 'hour-prices' in the course of day t. Based on given vectors $\mathbf{X}_{1:T} = \{X_1, X_2, \dots, X_T\} \subset R^{24}$, it is required to construct a forecast $\hat{X}_{T+1}$ for X_{T+1}. The forecasting algorithm [12] uses replacement of the vectors X_t by their low-dimensional features $Y_t = h(X_t) \in R^q$ (the LLE method is used; the value $q = 4$ is selected as an appropriate dimension of the features). Then the forecast $\hat{Y}_{T+1}$ for $Y_{T+1} = h(X_{T+1})$ based on the feature sample $\mathbf{Y}_{1:T} = h(\mathbf{X}_{1:T})$ in the reduced low-dimensional problem is constructed by using standard forecasting techniques. But then it is necessary to reconstruct the daily-price forecast $\hat{X}_{T+1}$ from the feature forecast $\hat{Y}_{T+1}$ which is the FOoS point in the general case.

A newly direction in the field of Machine Learning is meta-modeling in which data-based models (called meta-models [14] or surrogate models [15]) are constructed by learning on a set of input and output data prototypes obtained as a result of full-scale and/or computational experiments with some original complicated time-consuming models. As a rule, surrogate models have higher computational efficiency and can be used to replace original complicated models for further study (forecasting, optimization, etc.) [14, 15].

Input data which are original descriptions of objects under modeling can have high dimensionality, and the DR technique in meta-modeling is used for constructing reduced 'low-dimensional' surrogate models [16]. Thereafter, optimization or forecasting problems for the 'full-dimensional' model amounts to the corresponding reduced problems in the low-dimensional Feature space.

For example, meta-modeling is used in the wing shape optimization problem in aircraft designing [17], in which the DR is used to construct a low-dimensional wing airfoil parameterization [13]. In this problem the FOoS points appear as a result of solving optimization problems in the Feature space; thus, the reconstruction possibility is required in the DR.

However, the most of popular Manifold Embedding methods have a common drawback: they do not allow reconstructing high-dimensional points X from low-dimensional features h(X). Thus, it is necessary to formulate the ML problem in such a way that its solution does not have the above drawbacks. In other words, a corresponding ML procedure must reconstruct the unknown DM together with its low-dimensional parameterization from the sample.

We consider the ML problem called the Data Manifold Reconstruction problem, in which a low-dimensional representation of the DM allows accurate reconstruction of the DM [18, 19].

A strict definition is as follows: Given an input dataset $\mathbf{X}_n$ sampled from a q-dimensional DM $\mathbf{X}$ embedded in an ambient p-dimensional space R^p, $q < p$, and covered by a single chart, construct an **ML-**solution $\theta = (h, g)$ consisting of two inter-related mappings: an Embedding mapping h (2) and a Reconstruction mapping

$$g: \mathbf{Y} \subset R^q \to R^p,$$

which determine a reconstructed value $r_\theta(X) = g(h(X))$ as a result of successively applying the embedding and reconstruction mappings to a vector $X \in \mathbf{X}$. The solution must ensure the approximate equality

$$g(h(X)) \approx X \quad \text{for all} \quad X \in \mathbf{X}, \tag{3}$$

and the Reconstruction error $\delta_\theta(X) = |X - r_\theta(X)|$ is a measure of quality of the solution θ at a point $X \in \mathbf{X}$.

The Reconstruction mapping g must be defined not only on the feature sample $\mathbf{Y}_n$ (with an obvious reconstruction), but also on the FOoS features $y = h(X) \in \mathbf{Y} / \mathbf{Y}_n$ obtained by embedding the OoS points X.

The solution θ determines also a q-dimensional Reconstructed Manifold (RM)

$$\mathbf{X}_\theta = \{X = g(y) \in R^p: y \in \mathbf{Y}_\theta \subset R^q\} \tag{4}$$

embedded in R^p and parameterized by the chart g defined on the FS $\mathbf{Y}_\theta = h(\mathbf{X})$. The approximate equalities (3) can be considered as the Manifold proximity property

$$\mathbf{X}_\theta \approx \mathbf{X}, \tag{5}$$

meaning that the RM $\mathbf{X}_\theta = r_\theta(\mathbf{X})$ accurately reconstructs the DM $\mathbf{X}$ from the sample.

Note that the Data manifold reconstruction solution $\theta = (h, g)$ allows reconstructing the unknown DM $\mathbf{X}$ by the parameterized RM $\mathbf{X}_\theta$, whereas the Embedding Manifold solution h reconstructs a parameterization of the DM only.

From the statistical point of view, the defined Data manifold reconstruction problem may be considered as a Statistical Estimation Problem: Given a finite dataset randomly sampled from a smooth q-dimensional Data Manifold $\mathbf{X}$ covered by a single chart, estimate $\mathbf{X}$ by data-based q-dimensional manifold also covered (parameterized) by a single chart.

It is natural to evaluate the quality of the estimator $\mathbf{X}_\theta$ (4) (sample-based q-dimensional manifold in R^p also covered by a single chart) by the Hausdorff distance $H(\mathbf{X}_\theta, \mathbf{X})$ between the DM and RM [20]; the following relation between the qualities of the Data Manifold Reconstruction and Estimation problems takes a place:

$$H(\mathbf{X}_\theta, \mathbf{X}) \le \sup_{X \in \mathbf{X}} \delta_\theta(X).$$

Note that the defined Data Manifold Reconstruction problem differs from the Manifold approximation problem, in which an unknown manifold embedded in a high-dimensional ambient space must be approximated by some geometrical structure with close geometry, without any 'global parameterization' of the structure. For the latter problem, some solutions are known such as approximations by a simplicial complex

[21], by finitely many affine subspaces called 'flats' [22], tangential Delaunay complex [23], k-means and k-flats [24], and others. However, the Manifold approximation methods have a common drawback: they do not find a low-dimensional representation (parameterization) of the DM approximation; such parameterization is usually required in Machine Learning tasks with high-dimensional data.

There are some (though limited number of) methods for reconstruction of the DM $\mathbf{X}$ from the FS h($\mathbf{X}$). For a specific linear DM, the reconstruction can easily be made with the PCA. For a nonlinear DM, Auto-Encoder Neural Networks [3, 4, 5] determine both the embedding and reconstruction mappings. The LLE and LTSA methods also allow some reconstruction of the original vectors from their features.

5 Tangent Bundle Manifold Learning

The Reconstruction error $\delta_\theta(X)$ can be directly computed at sample points $X \in \mathbf{X}_n$; for an OoS point X it describes the generalization ability of the considered Data Manifold Reconstruction solution θ at a specific point X. Local lower and upper bounds are obtained for the maximum reconstruction error in a small neighborhood of an arbitrary point $X \in \mathbf{X}$ [19]; these bounds are defined in terms of the distance between the tangent spaces L(X) and $L_\theta(r_\theta(X))$ to the DM $\mathbf{X}$ and the RM $\mathbf{X}_\theta$ at the points X and $r_\theta(X)$, respectively. It follows from the bounds that the greater the distances between these tangent spaces, the lower the local generalization ability of the solution θ. Thus, it is natural to require that the MR-solution ensures not only Manifold proximity (5) but also Tangent proximity

$$L(X) \approx L_\theta(r_\theta(X)) \qquad \text{for all } X \in \mathbf{X} \tag{6}$$

between these tangent spaces in some selected metric on the Grassmann manifold Grass(p, q) consisting of all q-dimensional linear subspaces in R^p (the tangent spaces are treated as elements of the Grass(p, q)).

The requirement of the Tangent proximity for the Data Manifold Reconstruction solution arises also in various applications in which the MR is an intermediate step for Intelligent Data Analysis problem solution. For example, to ensure closeness between specific iterative optimization processes in the original and reduced design spaces, which are induced by the same optimization gradient-based method, it is necessary to guarantee accurate reconstruction of not only the DM (design space) $\mathbf{X}$ but also its tangent spaces. In Image Analysis, Data (Image) manifold may be very curved in an ambient space [25], and accurate reconstruction of the differential structure of the Image manifold (first of all, reconstruction of the tangent spaces) is required [26].

A statement of the extended Data Manifold Reconstruction problem, which includes an additional requirement of the tangent spaces proximity, has been proposed in [18, 19] and is described below.

The set $TB(\mathbf{X}) = \{(X, L(X)): X \in \mathbf{X}\}$ composed of points X of the manifold $\mathbf{X}$ equipped by tangent spaces L(X) at these points, is known in the Manifold theory as the Tangent Bundle of the manifold $\mathbf{X}$. Thus, accurate reconstruction of the DM $\mathbf{X}$ from the sample, which ensures accurate reconstruction of its tangent spaces too, can

be considered as reconstruction of the Tangent Bundle TB($\mathbf{X}$). Therefore, the amplification of the ML consisting in accurate reconstruction of the tangent bundle TB($\mathbf{X}$) from the sample $\mathbf{X}_n$ may be referred to as the Tangent Bundle Manifold Learning.

A strict definition of the TBML is as follows: Given dataset $\mathbf{X}_n$ sampled from a q-dimensional DM $\mathbf{X}$ embedded in an ambient p-dimensional Euclidean space R^p, $q < p$, construct TBML-solution $\theta = (h, g)$ which provides Tangent Bundle proximity consisting in the Manifold proximity (5) and the Tangent proximity (6), where $L_\theta(r_\theta(X)) = \mathrm{Span}(J_g(h(X)))$ is the tangent space to the RM $\mathbf{X}_\theta$ at a point $r_\theta(X)$ spanned by columns of the Jacobian $J_g(y)$ of the mapping $g(y)$ at a point $y = h(X) \in \mathbf{Y}_\theta$.

The TBML-solution θ determines the Reconstructed tangent bundle

$$\mathrm{RTB}_\theta(\mathbf{X}_\theta) = \{(g(y), \mathrm{Span}(J_g(y))): y \in \mathbf{Y}_\theta\} \tag{7}$$

of the RM $\mathbf{X}_\theta$, which is close to the TB($\mathbf{X}$), and the q-dimensional submanifold

$$\mathbf{L}_\theta = \{\mathrm{Span}(J_g(y)): y \in \mathbf{Y}_\theta\} \subset \mathrm{Grass}(p, q)$$

of the Grassmann manifold which reconstructs the Tangent Manifold

$$\mathbf{L} = \{L(X): X \in \mathbf{X}\} \subset \mathrm{Grass}(p, q).$$

The next section briefly describes the TBML-solution called the Grassmann & Stiefel Eigenmaps (GSE) algorithm [18, 19], which also gives new solutions for all the DR problems specified in Sections 2-4 above.

6 Tangent Bundle Manifold Learning Solution

The GSE algorithm consists of three successively performed steps: Tangent Manifold Learning, Manifold Embedding, and Tangent Bundle reconstruction.

In the Tangent Manifold Learning Step, a sample-based family $\mathbf{H} = \{H(X), X \in \mathbf{X}\}$ consisting of $p \times q$ matrices $H(X)$ smoothly depending on $X \in \mathbf{X}$ is constructed to meet the relations

$$L_H(X) \approx L(X) \quad \text{for all} \quad X \in \mathbf{X};$$

here $L_H(X)$ are q-dimensional linear spaces in R^p spanned by columns $\mathbf{H}^{(1)}(X)$, $\mathbf{H}^{(2)}(X)$, ... , $\mathbf{H}^{(q)}(X)$ of the matrices $H(X)$.

The family $\mathbf{H}$ is constructed in such a way as to provide the additional property: vector fields $\mathbf{H}^{(1)}(X)$, $\mathbf{H}^{(2)}(X)$, ... , $\mathbf{H}^{(q)}(X) \in L_\theta(r_\theta(X))$ must be potential and, therefore, meet the following relations

$$\nabla_{\mathbf{H}^{(i)}} \mathbf{H}^{(j)}(X) = \nabla_{\mathbf{H}^{(j)}} \mathbf{H}^{(i)}(X) \tag{8}$$

for $i, j = 1, 2, \ldots, q$; here $\nabla_{\mathbf{H}}$ denotes covariant differentiation with respect to the vector field $\mathbf{H}(X) \in L_\theta(r_\theta(X))$.

Let us briefly describe the Tangent Manifold Learning Step of the GSE. At first, the tangent space $L(X)$ for the points $X \in \mathbf{X}$ are estimated by the q-dimensional linear

space $L_{PCA}(X)$ which is a result of the PCA applied to sample points from an ε_n-ball in R^p centered at X; here $\varepsilon_n = O(n^{-1/(q+2)})$ is a small parameter.

The data-based kernel $K(X, X')$, $X', X \in \mathbf{X}$, is constructed as a product

$$K_E(X, X') \times K_G(X, X'),$$

where K_E is the Euclidean 'heat' kernel introduced in the LE algorithm [27] and

$$K_G(X, X') = K_{BC}(L_{PCA}(X), L_{PCA}(X'))$$

is the Binet–Cauchy kernel [28] on the Grass(p, q); this aggregate kernel reflects not only geometrical nearness between the points X and X′ but also nearness between the linear spaces $L_{PCA}(X)$ and $L_{PCA}(X')$, whence comes nearness between the tangent spaces L(X) and L(X′).

The set $\mathbf{H}_n$ consisting of explicitly written p×q matrices H_i that approximate the matrices $H(X_i)$, meet the constraints $Span(H_i) = L_{PCA}(X_i)$ and satisfy the conditions (8) written in a form of finite differences, is constructed to minimize the quadratic form

$$\Delta_{H,n}(\mathbf{H}_n) = \frac{1}{2}\sum_{i,j=1}^{n} K(X_i, X_j) \times \left\| H_i - H_j \right\|_F^2, \tag{9}$$

under the normalizing condition $\sum_{i=1}^{n} K(X_i) \times \left(H_i^T \times H_i\right) = I_q$ required to avoid a degenerate solution; here $K(X) = \sum_{j=1}^{n} K(X, X_j)$ and $K = \sum_{i=1}^{n} K(X_i)$.

Given $\mathbf{H}_n$, the p×q matrix H(X) for an arbitrary point $X \in \mathbf{X}$ is chosen to minimize the form $\Delta_H(H, X) = \sum_{j=1}^{n} K(X, X_j) \times \left\| H(X) - H_j \right\|_F^2$ under the specified linear conditions.

The exact solution of the minimizing problem (9) under the conditions (8) is obtained as a solution of specified generalized eigenvector problems. The matrix H(X) which minimizes the quadratic form $\Delta_H(H, X)$ is written in an explicit form.

This Step gives a new solution for the Tangent Manifold Learning problem of estimating the tangent spaces L(X) in the form of a smooth function of the point $X \in \mathbf{X}$, which was considered in some previous works. The matrices whose columns approximately span the tangent spaces were constructed using Artificial Neural Networks with one hidden layer [29] or Radial Basis Functions [30]. The constructed linear spaces $\{L_H(X_i)\}$ are the result of an alignment of the PCA-based linear spaces $\{L_{PCA}(X_i)\}$; a similar alignment problem was studied in the LTSA [31] with using a cost function which differs from our cost function (9).

The mappings h and g will be constructed in the next parts to provide the proximities

$$g(h(X)) \approx X \quad \text{and} \quad J_g(h(X)) \approx H(X), \tag{10}$$

whence comes the Tangent Bundle proximity (5), (6).

In the Manifold Embedding Step, given the family $\mathbf{H}$ already constructed, the Embedding mapping h(X) is constructed for $X \in \mathbf{X}$.

Taylor series expansions $g(y') - g(y) \approx J_g(y) \times (y' - y)$ for near points y, y′, under the desired equalities (10) for mappings h and g specified further, imply the equalities:

$$X' - X \approx H(X) \times (h(X') - h(X)) \tag{11}$$

for near points $X, X' \in \mathbf{X}$.

Under the family $\mathbf{H}$ already constructed, these approximate equalities can be considered as regression equations for the features h(X). First, consider equations (11) written for near sample points, and compute a preliminary vector set $\mathbf{Y}_n = \{y_1, y_2, \ldots, y_n\}$ as a standard least squares solution, which minimizes the weighted residual

$$\sum_{i,j=1}^{n} K(X_i, X_j) \times \left| X_j - X_i - H(x_i) \times (y_j - y_i) \right|^2$$

under the normalizing condition $y_1 + y_2 + \ldots + y_n = 0$.

Then, based on $\mathbf{Y}_n$, choose a value $y = h(X)$ for an arbitrary point $X \in \mathbf{X}$ by minimizing over y the weighted residual

$$\sum_{j=1}^{n} K(X, X_j) \times \left| X_j - X - H(X) \times (y_j - y) \right|^2.$$

Thus, under $\mathbf{Y}_n$, the value h(X) for an arbitrary point $X \in \mathbf{X}$ (including sample points) is written as

$$h(X) = h_{KNR}(X) + v^{-1}(X) \times Q_{PCA}^{T}(X) \times \tau(X),$$

here $v(X) = Q_{PCA}^{T}(X) \times H(X)$, $\tau(X) = \frac{1}{K(X)} \sum_{j=1}^{n} K(X, X_j) \times (X - X_j)$ and

$$h_{KNR}(X) = \frac{1}{K(X)} \sum_{j=1}^{n} K(X, X_j) \times y_j$$

is standard Kernel Non-parametric Regression estimator for h(X) based on the preliminary values $y_j \in \mathbf{Y}_n$ of the vector h(X) at the sample points.

The constrained mapping h determines the Feature space $\mathbf{Y}_\theta = h(\mathbf{X})$. This Step gives a new solution for the Manifold Embedding problem.

In the Tangent Bundle reconstruction step, given the family $\mathbf{H}$ and the mapping h already constructed, the mapping g is constructed to meet the proximities (3) and (6). This step gives a new solution for the Data Manifold Reconstruction.

The data-based kernel k(y, y′) on $\mathbf{Y}_\theta$ and the linear spaces $L^*(y) \in Grass(p, q)$ depending on $y \in \mathbf{Y}_\theta$ are constructed to provide the equalities

$$k(h(X), h(X')) \approx K(X, X')$$

and $L^*(h(X)) \approx L_{PCA}(X)$ for near points $X \in \mathbf{X}$ and $X' \in \mathbf{X}_n$.

The reconstruction function $\mathbf{g}(y)$ is constructed with using kernel nonparametric regression technique based on known values $X_i = \mathbf{g}(y_i)$ at the points $y_i = h(X_i)$ with taking into account the known values $J_g(y_i) = H(X_i)$, $i = 1, 2, \ldots, n$.

In the as asymptotic $n \to \infty$ and under an appropriate choice of the algorithm parameters, the rate in proximities (3) and (6) is $O(n^{-2/(q+2)})$ and $O(n^{-1/(q+2)})$, respectively [32]; the first rate coincides with the asymptotically minimax lower bound [20] for the Hausdorff distance between the DM $\mathbf{X}$ and RM $\mathbf{X}_\theta$. Thus, the RM $\mathbf{X}_\theta$ estimates the DM $\mathbf{X}$ with the optimal rate of convergence.

Acknowledgments. This work is partially supported by the Russian Foundation for Basic Research, research projects 13-01-12447 and 13-07-12111.

References

1. Bengio, Y., Courville, A., Vincent, P.: Representation Learning: A Review and New Perspectives. arXiv:1206.5538v2, 1–64 (2012)
2. Jollie, T.: Principal Component Analysis. Springer, New-York (2002)
3. Hecht-Nielsen, R.: Replicator neural networks for universal optimal source coding. Science 269, 1860–1863 (1995)
4. Kramer, M.: Nonlinear Principal Component Analysis using autoassociative neural networks. AIChE Journal 37(2), 233–243 (1991)
5. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006)
6. Schölkopf, B., Smola, A., Műller, K.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10(5), 1299–1319 (1998)
7. Cayton, L.: Algorithms for manifold learning. Univ. of California at San Diego (UCSD), Technical Report CS2008-0923, pp. 541–555. Citeseer (2005)
8. Huo, X., Ni, X., Smith, A.K.: Survey of Manifold-based Learning Methods. In: Liao, T.W., Triantaphyllou, E. (eds.) Recent Advances in Data Mining of Enterprise Data, pp. 691–745. World Scientific, Singapore (2007)
9. Ma, Y., Fu, Y. (eds.): Manifold Learning Theory and Applications. CRC Press, London (2011)
10. Bengio, Y., Delalleau, O., Le Roux, N., Paiement, J.-F., Vincent, P., Ouimet, M.: Learning Eigenfunctions Link Spectral Embedding and Kernel PCA. Neural Computation 16(10), 2197–2219 (2004)
11. Lee, J.A., Verleysen, M.: Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing 72(7-9), 1431–1443 (2009)
12. Chen, J., Deng, S.-J., Huo, X.: Electricity price curve modeling and forecasting by manifold learning. IEEE Transaction on Power Systems 23(3), 877–888 (2008)
13. Bernstein, A.V., Burnaev, E.V., Chernova, S.S., Zhu, F., Qin, N.: Comparison of Three Geometric Parameterization methods and Their Effect on Aerodynamic Optimization. In: Proceedings of International Conference on Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and Societal Problems, Eurogen 2011, Capua, Italy, September 14-16 (2011)
14. Gary Wang, G., Shan, S.: Review of Metamodeling Techniques in Support of Engineering Design Optimization. J. Mech. Des. 129(3), 370–381 (2007)
15. Forrester, A.I.J., Sobester, A., Keane, A.J.: Engineering Design via Surrogate Modelling. A Practical Guide. Wiley, New-York (2008)
16. Kuleshov, A.P., Bernstein, A.V.: Cognitive Technologies in Adaptive Models of Complex Plants. Information Control Problems in Manufacturing 13(1), 1441–1452 (2009)

17. Bernstein, A., Kuleshov, A., Sviridenko, Y., Vyshinsky, V.: Fast Aerodynamic Model for Design Technology. In: Proceedings of West-East High Speed Flow Field Conference, WEHSFF-2007. IMM RAS, Moscow (2007), http://wehsff.imamod.ru/pages/s7.html
18. Bernstein, A.V., Kuleshov, A.P.: Tangent Bundle Manifold Learning via Grassmann & Stiefel Eigenmaps. In arXiv preprint: arXiv:1212.6031v1 [cs.LG], pp. 1–25 (December 2012)
19. Bernstein, A.V., Kuleshov, A.P.: Manifold Learning: generalizing ability and tangent proximity. International Journal of Software and Informatics 7(3), 359–390 (2013)
20. Genovese, C.R., Perone-Pacifico, M., Verdinelli, I., Wasserman, L.: Minimax Manifold Estimation. Journal Machine Learning Research 13, 1263–1291 (2012)
21. Freedman, D.: Efficient simplicial reconstructions of manifold from their samples. IEEE Transaction on Pattern Analysis and Machine Intelligence 24(10), 1349–1357 (2002)
22. Karygianni, Sofia, Frossard, Pascal. Tangent-based manifold approximation with locally linear models. In: arXiv:1211.1893v1 [cs.LG] (November 6, 2012)
23. Boissonnat, J.-D., Ghosh, A.: Manifold reconstruction using tangential Delaunay complexes. Discrete & Computational Geometry 51(1), 221–267 (2014)
24. Canas, G.D., Poggio, T., Rosasco, L.A.: Learning Manifolds with K-Means and K-Flats. In: Advances in Neural Information Processing Systems, NIPS (2012)
25. Pless, R., Souvenir, R.: A Survey of Manifold Learning for Images. IPSJ Transactions on Computer Vision and Applications 1, 83–94 (2009)
26. Kuleshov, A.P., Bernstein, A.V.: Tangent Bundle Manifold Learning for Image Analysis. In: Vuksanovic, B., Verikas, A., Zhou, J. (eds.) Proceedings of SPIE, Sixth International Conference on Machine Vision, ICMV 2013, London, The United Kingdom, November 16-17, vol. 9067, pp. 201–205 (2013)
27. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15, 1373–1396 (2003)
28. Wolf, L., Shashua, A.: Learning over sets using kernel principal angles. J. Mach. Learn. Res. 4, 913–931 (2003)
29. Bengio, Y., Monperrus, M.: Non-local manifold tangent learning. In: Saul, L., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems, vol. 17, pp. 129–136. MIT Press, Cambridge (2005)
30. Dollár, P., Rabaud, V., Belongie, S.: Learning to Traverse Image Manifolds. In: Schölkopf, B., Platt, J.C., Hoffman, T. (eds.) Advances in Neural Information Processing Systems, 19, pp. 361–368. MIT Press, Cambridge (2006)
31. Zhang, Z., Zha, H.: Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment. SIAM Journal on Scientific Computing 26(1), 313–338 (2005)
32. Kuleshov, A., Bernstein, A., Yanovich, Y.: Asymptotically optimal method in Manifold estimation. In: Márkus, L., Prokaj, V. (eds.) Abstracts of the XXIX-th European Meeting of Statisticians, Budapest, July 20-25, p. 325 (2013)