Int J Software Informatics, Volume 7, Issue 3 (2013), pp. 359–390 Tel: +86-10-62661040
International Journal of Software and Informatics, ISSN 1673-7288 http://www.ijsi.org

Email: ijsi@iscas.ac.cn, ijsi2007@gmail.com

Manifold Learning: Generalization Ability and Tangent Proximity

Alexander V. Bernstein[1,2,3] and Alexander P. Kuleshov[1,2]

[1] (Institute for Information Transmission Problems, Russian Academy of Sciences,
Moscow 127994, Russia)

[2] (Department of Technologies of Complex Systems Modeling,
National Research University Higher School of Economics, Moscow 109028, Russia)

[3] (PreMoLab, Moscow Institute of Physics and Technology (State University),
Moscow 115184, Russia)

Abstract One of the ultimate goals of Manifold Learning (ML) is to reconstruct an unknown nonlinear low-dimensional Data Manifold (DM) embedded in a high-dimensional observation space from a given set of data points sampled from the manifold. We derive asymptotic expansion and local lower and upper bounds for the maximum reconstruction error in a small neighborhood of an arbitrary point. The expansion and bounds are defined in terms of the distance between tangent spaces to the original DM and the Reconstructed Manifold (RM) at the selected point and its reconstructed value, respectively. We propose an amplification of the ML, called Tangent Bundle ML, in which proximity is required not only between the DM and RM but also between their tangent spaces. We present a new geometrically motivated Grassman&Stiefel Eigenmaps algorithm that solves this problem and gives a new solution for the ML also.

Key words: dimensionality reduction; manifold learning; generalization ability; tangent spaces; tangent bundle manifold learning; Grassmann manifold; Stiefel manifold

Bernstein AV, Kuleshov AP. Manifold learning: generalization ability and tangent proximity. ***Int J Software Informatics*****, Vol.7, No.3 (2013): 359–390. http://www.ijsi.org/1673-7288/7/i165.htm**

1 Introduction

The goal of Dimensionality Reduction (DR) is extracting low-dimensional structure from high-dimensional data. There exist a number of methods (techniques) for the DR. Linear DR is well known and uses such techniques as Principal Component Analysis[34] (PCA) and classical metric Multidimensional Scaling[18] (MDS). Various Non-linear DR techniques are based on Auto-Encoder neural networks[30, 31, 38, 49], Self-organizing Maps[37], Topology representing networks[48], Diffusion Maps[41], Kernel PCA[55], and others.

This work is partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF Government grant, ag. 11.G34.31.0073; and RFBR, research projects 13-01-12447 and 13-07-12111.
Corresponding author: Alexander V. Bernstein, Email: a.bernstein@mail.ru
Received 2012-11-30; Revised 2013-06-03; Accepted 2013-08-31.

A newly emerging direction in the fields of the DR, which has been a subject of intensive research over the last decades, consists in constructing a family of DR-algorithms based on studying local structure of a given sampled dataset: Locally Linear Embedding[52] (LLE); Laplacian Eigenmaps[2] (LE); Hessian Eigenmaps[21] (HE); ISOmetric MAPing[63] (ISOMAP); Maximum Variance Unfolding[68] (MVU); Manifold charting[10]; Local Tangent Space Alignment[72,73] (LTSA), and others. Some of these algorithms (LLE, LE, ISOMAP, MVU) can be considered in the same framework, based on the Kernel PCA[55] applied to various data-based kernels[3, 4, 53, 54].

The DR problems are formulated in various ways, and we will give a few different DR formalizations, including a new formalization proposed in the paper. There is no generally accepted terminology in the DR; thus, some terms introduced below can be different from those used in some other works.

1.1 Dimensionality reduction as Data Space (Manifold) Embedding

One formalization of the DR, which is used in most of the above listed papers and can be referred to as the **Sample Embedding** problem, is as follows: Given an input dataset (sample)

$$\boldsymbol{X}_n = \{X_1, X_2, \ldots, X_n\} \subset \boldsymbol{X} \subset R^p, \tag{1}$$

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

$$h_{(n)} : \boldsymbol{X}_n \subset R^p \to \boldsymbol{h}_n = h_{(n)}(\boldsymbol{X}_n) = \{h_1, h_2, \ldots, h_n\} \subset R^q \tag{2}$$

of the sample $\boldsymbol{X}_n$ to an q-dimensional dataset $\boldsymbol{h}_n, q < p$, which *faithfully represents* the high-dimensional sample $\boldsymbol{X}_n$ while inheriting certain subject-driven data properties like preserving the local data geometry (LLE, LTSA), proximity relations (LE, HE), geodesic distances (ISOMAP), angles (Conformal Eigenmaps[56]; Conformal and Landmark ISOMAP[60]), etc.

The term *'faithfully represents'* is not formalized in general, and in various DR methods it is different due to choosing some optimized cost function $L_{(n)}(h_1, h_2, \ldots, h_n|\boldsymbol{X}_n)$ which defines an 'evaluation measure' for the DR and reflects desired properties of the n-point Embedding mapping $h_{(n)}$ (2). As is pointed out in Ref. [13], a general view on the DR can be based on a 'concept of cost functions'. For example, the cost function

$$L_{(n)}\left(h_1, h_2, \ldots, h_n \left|\mathbf{X}_n\right.\right) = \sum_{i,j} \left(\rho(X_i, X_j) - \|h_i - h_j\|\right)^2,$$

is considered in the classical MDS, here ρ is a chosen metric in the DS $\boldsymbol{X}$. Note that the MDS and PCA methods are equivalent when ρ is the Euclidean metric in R^p.

An extehsion of the Sample Embedding, which can be referred to as the **Data Space Embedding** problem, is as follows: Given an input dataset (sample) $\boldsymbol{X}_n$ (1) from the DS $\boldsymbol{X}$, construct a low-dimensional parameterization of the DS which produces an **Embedding mapping**

$$h : \boldsymbol{X} \subset R^p \to \boldsymbol{Y} = h(\boldsymbol{X}) \subset R^q \tag{3}$$

from the DS to a Reconstructed Coordinate (Feature) Space $\boldsymbol{Y} = h(\boldsymbol{X})$, which preserves specific properties of the DS. Note that the mapping h must also be defined on new **Out-of-Sample** (OoS) points $X_{new} \in \boldsymbol{X}/\boldsymbol{X}_n$.

Based on some solution $h_{(n)}$ for the Sample Embedding problem from sample $\boldsymbol{X}_n$, an emdedding mapping for new OoS points $X_{new} \in \boldsymbol{X}/\boldsymbol{X}_n$ could certainly be constructed as $h_{(n+1)}(X_{new})$ with regard to Sample Embedding $h_{(n+1)}$ applied to the dataset $\{\boldsymbol{X}_n \cup X_{new}\}$; however, $h_{(n+1)}(\boldsymbol{X}_n)$ can not coincide in the general case with the embedding $\boldsymbol{h}_n = h_{(n)}(\boldsymbol{X}_n)$ obtained previously for the initial sample $\boldsymbol{X}_n$. An '**OoS extension**' for the algorithms LLE, LE, ISOMAP and MVU, which are based on the Kernel PCA approach, has been found in Ref. [4] with using Nyström's eigendecomposition technique[14,53]. The Cost functions concept[13] allows constructing an Embedding mapping for the OoS points; other OoS techniques are proposed in Ref. [24,62], etc.

The definition of the Data Space Embedding problem uses values of the function h (3) for OoS points; thus, we must define a **Data Model** describing the DS, and a **Sampling Model** offering a way for extracting both the sample $\boldsymbol{X}_n$ and OoS points from the DS $\boldsymbol{X}$.

The most popular models in the DR are **Manifold Data Models**[12,15,20,26,35,46,47,64,72], in which the DS $\boldsymbol{X}$ is a q-dimensional manifold embedded in p-dimensional Euclidean space $R^p, q < p$, and referred to as the **Data Manifold** (DM). A motivation of such model consists in the follofing empirical fact: as a rule, the real-world data presented in high-dimensional spaces is likely to concentrate in a vicinity of a non-linear submanifold of much lower dimensionality[15]; this assumption is referred to as **Manifold hypothesis** or **Manifold assumption**. The Data Space Embedding problem with the Manifold Data Model can be referred to as the **Manifold Embedding** problem.

In most studies, DM is modeled using a single coordinate chart:

$$X = \{X = f(b) \in R^p : b \in \boldsymbol{B} \subset R^q\} \subset R^p, \tag{4}$$

and it is supposed that $\boldsymbol{X}$ is a well-behaved manifold: a coordinate chart f is a diffeomorphism from an open subset $\boldsymbol{B}$ in R^q to DM $\boldsymbol{X} = f(\boldsymbol{B})$ with a differentiable inverse map, and the manifold $\boldsymbol{X}$ has no self-intersections. The set $\boldsymbol{B}$ will referred to as the **Coordinate space**,

The Sampling Model is typically defined as a probability measure μ on the σ-algebra of measurable subsets of the DM $\boldsymbol{X}$ whose support Supp(μ) coincides with $\boldsymbol{X}$. In accordance with this model, the dataset $\boldsymbol{X}_n$ (1) and OoS points $X \in \boldsymbol{X}/\boldsymbol{X}_n$ are selected from the DM $\boldsymbol{X}$ independently of each other according to the probability measure μ.

1.2 Manifold reconstruction in dimensionality reduction

Manifold Embedding is usually a first step in various Intelligent Data Analysis problems (classification, clustering, etc.): a reduced q-dimensional vector $y = h(X)$ is used in their procedures instead of an initial p-dimensional vector X.

If the Embedding mapping h (3) in the Manifold Embedding preserves only specific properties of high-dimensional data, then substantial data losses are possible when using a reduced q-dimensional vector $y = h(X)$ instead of the initial p-dimensional vector X. As is pointed out in Ref. [43,44], one objective of the DR is to preserve as much available information contained in the sample as possible. Under this approach, the term *'faithfully represents'* is understood as preserving of such information, and this means the possibility of reconstructing high-dimensional points X from low-dimensional 'embedded' points $h(X)$. This possibility can be considered as a valid evaluation measure for the DR procedures[43,44].

To prevent these losses in the general case, an embedding mapping must provide ability for reconstructing the initial vector $X \in \boldsymbol{X}$ from a vector $y = h(X)$ with small reconstruction error. Thus, it is necessary to construct a **Reconstruction mapping**

$$g : \boldsymbol{Y}_g \subset R^q \to R^p \tag{5}$$

defined on a domain $\boldsymbol{Y}_g \supset \boldsymbol{Y} = h(X)$, which determines a reconstructed value

$$X^* = (g \cdot h)(X) \equiv g(h(X)) \tag{6}$$

as a result of successively applying the embedding and reconstruction mappings to a vector $X \in \boldsymbol{X}$. The Reconstruction mapping g must be defined not only on the Embedded sample $\boldsymbol{Y}_n = h(\boldsymbol{X}_n)$ (with an obvious reconstruction), but also on 'OoS' embedded points $y = h(X) \in \boldsymbol{Y} / \boldsymbol{Y}_n$ obtained by embedding the OoS points $X \in \boldsymbol{X}/\boldsymbol{X}_n$.

Problems in which a reconstruction mapping is required arise in numerous applications.

Example 1 concerns a Wing shape optimization problem, which is one of important problems in aircraft designing. Design variables include a number of high-dimensional vectors X of dimension p which are detailed descriptions of wing airfoils. In practical applications, the dimension p varies in the range from 50 to 200; specific value of p is selected depending on a required accuracy of airfoil description.

Low-dimensional airfoil parameterization[61] is usually constructed in order to reduce the number of design variables, and Dimensionality Reduction technique is one of highly powerful methods for such parameterization[6,8]. This technique constructs a low-dimensional parameterization (Embedding mapping h (3)) of a given airfoil based on a known dataset consisting of high-dimensional descriptions of airfoils-prototypes and allows to describe airfoils of aircraft wings by low-dimensional vectors whose dimension q varies in the range from 5 to 10.

The constructed airfoil parameterization determines a low-dimensional design space, and a wing shape optimization problem is reduced to optimizing some functional on this space. If low-dimensional airfoil descripion y^* is a result of some optimization procedure in the constructed design space, then it is required to reconstruct a detailed description $X^* = g(y^*)$ (6) of the 'optimal' wing airfoil for the 'OoS' value y^*. □

Example 2, described in Ref. [16], is devoted to *Processing of an Electricity price curve.* Electricity 'daily' prices are described by a multidimensional time series (electricity price curve) $X_t = (X_{t1}, X_{t2}, \ldots, X_{t,24})^T \in R^{24}, t = 1, 2, \ldots, T$, consisting of 'hour-prices' in the course of day t. Based on $X_1, X_2, \ldots, X_T$, it is necessary to construct a forecast X^* for X_{T+1}. Both Embedding and Reconstruction mappings are used in the Forecasting algorithm[16]:

- LLE[52] is used for solving the Manifold Embedding problem on the basis of the given 'daily-prices' vectors $\{X_1, X_2, \ldots, X_T\} \subset R^{24}$. Denote by h the constructed LLE-based solution and by $\{y_t = h(X_t) = (y_{t1}, y_{t2}, \ldots, y_{tq})^T \in R^q, t = 1, 2, \ldots, T\}$ the embedding results ($q = 4$ is selected in Ref. [16] as appropriate dimension of the embeddings).

- Based on a one-dimensional time series $\{y_{tk}, t = 1, 2, \ldots, T\}$, by using standard forecasting technique, a forecast Y_k for $y_{T+1,k}$ is constructed for $k = 1, 2, \ldots, q$.

- Based on the constructed q-dimensional vector $Y^* = (Y_1, Y_2, \ldots, Y_q)^T$, a forecast $X^* = g(Y^*)$ for X_{T+1} is constructed with using LLE-reconstruction technique[53].

As is pointed out in Ref. [16] (citations): 'low-dimensional coordinates to the high-dimensional space is a necessary step for forecasting'; 'the reconstruction of high-dimensional forecasted price curves from low-dimensional prediction is a significant step for forecasting'; 'reconstruction accuracy is critical for the application of manifold learning in the prediction.' □

There are some (though limited number of) methods for reconstruction the DS from the Reconstructed Coordinate Space $\boldsymbol{Y} = h(\boldsymbol{X})$. For a specific linear DM, the reconstruction can be easily made with the PCA. For a nonlinear DM, the constructed sample-based Auto-Encoder Neural Network determines both the embedding and reconstruction mappings. LLE reconstruction, which is done in the same manner as LLE, has been introduced in Ref. [53]. LTSA reconstruction, an interpolation-like reconstruction, and nonparametric regression reconstruction have been proposed in Ref. [72].

The Reconstruction mapping g (5) determines a **Reconstruction error**

$$\delta(X) = \|X - g(h(X))\| \tag{7}$$

at a point $X \in \boldsymbol{X}$. The Reconstruction error can be chosen as an evaluation measure in the DR: small value of $\delta(X)$ means that $h(X)$ well preserves information contained in X. Thus, the error $\delta(X)$ may be considered as an 'universal quality criterion'[43,44] in the DR.

1.3 Dimensionality reduction as Manifold Learning

General DR with Manifold Data Model is often referred to as the **Manifold Learning** (ML) problems. In this paper, by ML we will mean the DR in which the term *'faithfully represents'* has the following specified strong meaning: a low dimensional representation of the DM must preserve as much information contained in high-dimensional data as possible. Thus, a specified evaluation measure of DR quality has to reflect preservation of this information, and the Reconstruction error may be chosen as such measure.

A strict definition of the ML is as follows: Given an input dataset $\boldsymbol{X}_n$ (1) sampled from a q-dimensional Data Manifold $\boldsymbol{X}$ (4) covered by a single chart, construct an **ML-solution** $\theta = (h, g)$ consisting of an Embedding mapping h (3) defined on a domain of definition $\boldsymbol{X}_h \supset \boldsymbol{X}$ and a Reconstruction mapping g (5) defined on a domain of definition $\boldsymbol{Y}_g \supset \boldsymbol{Y}_h = \boldsymbol{h}(\boldsymbol{X}_h)$, which ensures the approximate equality

$$X \approx g(h(X)) \quad \text{for all} \quad X \in \boldsymbol{X}. \tag{8}$$

In this definition, the Reconstruction error $\delta(X)$ (7), which equals to Euclidean norm of the residual vector

$$\Delta(X) = X - g(h(X)),$$

is a measure of quality of the ML solution at a point $X \in \boldsymbol{X}$.

The mapping g determines a q-dimensional manifold

$$\boldsymbol{X}_g = \{X = g(y) \in R^p : y \in \boldsymbol{Y}_g \subset R^q\} \tag{9}$$

embedded in R^p and covered (parameterized) by a single coordinate chart g defined on its domain of definition $\boldsymbol{Y}_g$. A constriction of the domain $\boldsymbol{Y}_g$ on the set

$$\boldsymbol{Y}_\theta = h(\boldsymbol{X}) \subset \boldsymbol{Y}_g \tag{10}$$

determines a q-dimensional **Reconstructed Manifold** (RM)

$$\boldsymbol{X}_\theta = \{X = g(y) \in R^p : y \in \boldsymbol{Y}_\theta \subset R^q\} \tag{11}$$

embedded in R^p and parameterized by the chart g defined on $\boldsymbol{Y}_\theta$. The coordinate space $\boldsymbol{Y}_\theta$ of the RM $\boldsymbol{X}_\theta$ will be called the **Reconstructed Coordinate Space**.

The solution $\theta = (h, g)$ determines the mapping

$$r_\theta(X) = g(h(X)) \tag{12}$$

from the DM $\boldsymbol{X}$ into the RM $\boldsymbol{X}_\theta = r_\theta(\boldsymbol{X})$, and the approximate equalities (8) can be considered as **Manifold proximity** property

$$\boldsymbol{X} \approx \boldsymbol{X}_\theta, \tag{13}$$

meaning that the RM $\boldsymbol{X}_\theta$ accurately reconstructs (estimates) the DM $\boldsymbol{X}$ from the sample $\boldsymbol{X}_n$. Thus, the ML solution $\theta = (h, g)$ allows reconstructing the unknown DM $\boldsymbol{X}$ by the parameterized RM $\boldsymbol{X}_\theta$, whereas the Embedding Manifold solution h (3) reconstructs a parameterization of the DM only.

From *Statistical point of view*, the defined ML problem may be considered as a Statistical Estimation Problem: there is an unknown object $\boldsymbol{X}$ (4) (smooth q-dimensional Data Manifold in R^p covered by a single chart) and a finite dataset $\boldsymbol{X}_n$ randomly sampled from $\boldsymbol{X}$. Based on the sample, it is required to construct an estimator $\boldsymbol{X}_\theta$ (11) (also a q-dimensional manifold in R^p covered by a single chart and constructed from the sample) for $\boldsymbol{X}$; this estimator is determined by a pair $\theta = (h, g)$ of mappings (3), (5). Quality of a solution $\theta = (h, g)$ is defined as accuracy in the approximated equality (13), and the Reconstruction error $\delta(X)$ (7) is a quality measure at a specific point $X \in \boldsymbol{X}$.

The defined ML Problem differs from the *Manifold approximation problem*, also called *Manifold reconstruction*, which is as follows: Given a finite dataset randomly sampled from an unknown manifold, represent the manifold geometry by some geometrical structure in the original ambient space R^p, without any 'global parameterization' and any mapping from the initial manifold to the approximated structure. For the latter problem, some solutions are known such as approximations by a simplicial complex[23] or by finitely many affine subspaces called 'flats'[36].

1.4 Short review of the paper results

The Reconstruction error $\delta(X)$ (7) can be directly computed for sample points $X \in \boldsymbol{X}_n$, and for OoS point $X \in \boldsymbol{X} \backslash \boldsymbol{X}_n$ it describes the **generalization ability** of the considered ML solution (h, g) at a specific point X. In Section 2, asymptotic expansion and local lower and upper bounds are obtained for the maximum reconstruction error in a small neighborhood of an arbitrary point $X \in \boldsymbol{X}$. The expansion and lower and upper bounds are defined in terms of the distance between tangent spaces to the DM $\boldsymbol{X}$ (4) and the RM $\boldsymbol{X}_\theta$ (11) at the considered point $X \in \boldsymbol{X}$ and the reconstructed point $r_\theta(X) \in \boldsymbol{X}_\theta$ (12), respectively. It follows from these results that the greater the distances between these tangent spaces, the lower the local generalization ability of the considered ML procedure.

Thus, it is natural to require that the ML procedure (h, g) ensures not only proximity (8) between the points $X \in \boldsymbol{X}$ and their reconstructed values $r_\theta(X)$ (12) but also proximity between the corresponding tangent spaces. A statement of the extended ML problem, which may be referred to as the **Tangent Bundle Manifold Learning** (TBML) problem and includes a requirement of tangent spaces proximity, is proposed in Section 3.

A solution of the TBML based on the proposed **Grassmann&Stiefel Eigenmaps** (GSE) approach, which also gives a new solution for the ML, is described in Section 4. Results of performed comparative numerical experiments are presented in Section 5.

1.5 List of mathematical notations

This subsection contains a list of mathematical notations used throughout the paper:

$\boldsymbol{X}$ – Data Manifold; q-dimensional manifold embedded in p-dimensional ambient space R^p;

$\boldsymbol{X}_n = \{X_1, X_2, \dots, X_n\} \subset \boldsymbol{X}$ – finite dataset (sample) randomly sampled from the $\boldsymbol{X}$;

h – Embedding mapping from a domain of definition $\boldsymbol{X}_h \supset \boldsymbol{X}$ to R^q;

$\boldsymbol{Y}_n = h(\boldsymbol{X}_n) = \{y_1, y_2, \dots, y_n\}$ – the result of sample's embedding;

g – Reconstruction mapping from a domain of definition $\boldsymbol{Y}_g \supset \boldsymbol{Y}_h = h(X_h)$ to R^p;

$\theta = (h, g)$ – ML solution;

$\boldsymbol{Y}_\theta = h(\boldsymbol{X})$ – low dimensional image of Data Manifold; Reconstructed Coordinate Space;

$\boldsymbol{X}_\theta = \{X = g(y) \in R^p : y \in \boldsymbol{Y}_\theta \subset R^q\}$ – Reconstructed Manifold;

$r_\theta(X) = g(h(X)) \in \boldsymbol{X}_\theta = r_\theta(\boldsymbol{X})$ – reconstructed value of the point $X \in \boldsymbol{X}$;

$\delta_\theta(X) = \|X - g(h(X))\|$ – Reconstructed error;

L(X) – tangent space to the Data Manifold at the point $X \in \boldsymbol{X}$;

$L_\theta(r_\theta(X))$ – tangent space to the Reconstructed Manifold at the point $r_\theta(X) \in \boldsymbol{X}_\theta$.

2 Local Generalization Ability in Manifold Learning

2.1 Basic concepts and notations

Let $X_0 \in \boldsymbol{X}$ be some selected point, and let

$$\delta_\theta(X_0, \varepsilon) = \max\{\delta_\theta(X) : X \in U_\varepsilon(X_0)\} \tag{14}$$

be the maximum reconstruction error in the ε-neighborhood

$$U_\varepsilon(X_0) = \{X \in \boldsymbol{X} : \|X - X_0\| \leqslant \varepsilon\}$$

of the point X_0. The quantity $\delta_\theta(X_0, \varepsilon)$ (14) characterizes the local generalization ability of the procedure θ in the ε-neighborhood of the point X_0.

As before, we will assume in the paper that $\boldsymbol{X}$ is a well-behaved manifold and has also the following additional properties: Jacobian ($p \times q$ matrix) $J_f(b)$ of the diffeomorphism f has rank q for all $b \in \boldsymbol{B}$, and there are positive constants C, C', C'' such that if two arbitrary points $X = f(b)$ and $X' = f(b')$ from $\boldsymbol{X}$ satisfy the condition $\|X - X'\| < C$ then the inequalities

$$C' \times \|b - b'\| \leqslant \|X - X'\| \leqslant C'' \times \|b - b'\|$$

hold true. Therefore, the DM $\boldsymbol{X}$ has a tube $\text{Tube}_\varepsilon(\boldsymbol{X})$ of some positive radius ε (that is, the points from an ε-neighborhood of $\boldsymbol{X}$ have a single projection onto $\boldsymbol{X}$), whence comes that $\boldsymbol{X}$ has no self-intersections.

Let $\theta = (h, g)$ be some ML-solution. If the DM $\boldsymbol{X}$ lies in a positive radius tube Tube $(\boldsymbol{X}_g)$ of the manifold $\boldsymbol{X}_g$ (9), one can consider a new solution $\theta(g) = (h_g, g)$, where

$$h_g(X) = \arg\min_y\{\|g(y) - X\| : y \in \boldsymbol{Y}_g\}; \tag{15}$$

i.e., the point $r_{\theta(g)}(X) = g(h_g(X)) \in \boldsymbol{X}_g$ is the projection of the point $X \in \boldsymbol{X} \subset$ Tube $(\boldsymbol{X}_g)$ onto the manifold $\boldsymbol{X}_g$ (9). The solution θ(g) determines also the Reconstructed Manifold

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

with the Reconstructed Coordinate Space $\boldsymbol{Y}_{\theta(g)} = h_g(\boldsymbol{X})$.

By definition, we have the inequality

$$\|X - r_{\theta(g)}(X)\| \leqslant \|X - r_\theta(X)\| \quad \text{for all} \quad X \in \boldsymbol{X}.$$

Hence, the new solution θ(g) has smaller reconstruction error and thus improves the initial solution θ. Because of this, in this section we will study the generalization ability of the improved solution θ(g).

To formulate the obtained results, introduce some notations. Let $p \times q$ matrices $J_f(b)$ and $J_g(y)$ be Jacobians of the mappings f (4) and g (5), respectively. The q-dimensional linear subspaces in R^p

$$\begin{aligned} L(X) &= \text{ Span}(J_f(b)), \\ L_g(y) &= \text{ Span}(J_g(y)), \end{aligned} \tag{16}$$

which are spanned by columns of these Jacobians, are the tangent spaces to the manifolds $\boldsymbol{X}$ and $\boldsymbol{X}_{\theta(g)}$ at the points $X = f(b) \in \boldsymbol{X}$ and $g(y) \in \boldsymbol{X}_{\theta(g)}$, respectively; here $b \in \boldsymbol{B}$ and $y \in \boldsymbol{Y}_{\theta(g)}$.

Let $X_0 = f(b_0) \in \boldsymbol{X}$ be an arbitrary point, $b_0 \in \boldsymbol{B}$, and let $y_0 = h_g(X_0)$. Denote by $\zeta_1, \zeta_2, \ldots, \zeta_q$ the principal angles[25,32,33] between the subspaces $L(X_0)$ and $L_g(y_0)$ arranged in ascending order:

$$0 \leqslant \zeta_1 \leqslant \zeta_2 \leqslant \ldots \leqslant \zeta_q \leqslant \pi/2, \tag{17}$$

and denote $\{t_{f,1}, t_{f,2}, \ldots, t_{f,q}\}$ and $\{t_{g,1}, t_{g,2}, \ldots, t_{g,q}\}$ the principal vectors in the subspaces $L(X_0)$ and $L_g(y_0)$, respectively. These vectors determine orthonormal bases for these subspaces and satisfy the relations

$$(t_{f,i}, t_{g,j}) = \delta_{ij} \times \cos(\zeta_j), \quad i, j = 1, 2, \ldots, q, \tag{18}$$

hereinafter $(\cdot,\cdot)$ denotes the scalar product in R^p. Let

$$t_{f,g,j} = t_{f,j} - t_{g,j} \times \cos(\zeta_j) \equiv \pi^{\perp}(y_0) \times t_{f,j}, \quad j = 1, 2, \ldots, q, \tag{19}$$

be projections of the principal vectors $\{t_{f,1}, t_{f,2}, \ldots, t_{f,q}\}$ onto the subspace $(L_g(y_0))^{\perp}$. It follows from (18), (19) that these vectors are orthogonal and satisfy the relations

$$(t_{f,g,i}, t_{f,g,j}) = \delta_{ij} \times \|t_{f,g,j}\|^2 = \delta_{ij} \times \sin^2(\zeta_j), \quad i, j = 1, 2, \ldots, q. \tag{20}$$

Denote by $L^* \subseteq (L_g(y_0))^{\perp}$ a linear space spanned by the vectors $\{t_{f,g,1}, t_{f,g,2}, \ldots, t_{f,g,q}\}$; that is, the linear space L^* is the orthogonal complement to the projection of the tangent space $L(X_0)$ onto the tangent space $L_g(y_0)$. Let $q^* = q - k^*$ be a dimension of the L^*, where

$$k^* = \max\{k : \zeta_j = 0, \quad j = 1, 2, \ldots, k\}$$

under $\zeta_1 = 0$; otherwise put $k^* = 0$. Note that if $k^* = q$ (that is, $\zeta_q = 0$ and the subspaces $L(X_0)$ and $L_g(y_0)$ coincide), then $q^* = k^* - q = 0$ and L^* is a degenerate linear space. The unit vectors

$$e_j = t_{f,g,k^*+j} \times \sin^{-1}(\zeta_{k^*+j}), \quad j = 1, 2, \ldots, q^*,$$

form an orthonormal basis in L^*, see (20).

Denote by $A_j = (X_0 - r_{\theta(g)}(X_0), e_j), j = 1, 2, \ldots, q^*$, the components of the projection

$$A = \pi_{L^*}(\Delta(X_0)) = (A_1, A_2, \ldots, A_{q^*})^T \tag{21}$$

of the resudial vector $\Delta(X_0)$ onto the linear space L^*.

2.2 The theorem on the generalization ability

Theorem 1. Let h (3) and g (5) be smooth mappings whose Jacobians have rank q. Let the DM $\boldsymbol{X}$ lies in a tube Tube $(\boldsymbol{X}_g)$ of the RM $\boldsymbol{X}_g$ of a positive radius.

Then we have the following asymptotic expansion of the local maximum reconstruction error $\delta_{\theta(g)}(X_0,\,\varepsilon)$:

$$\delta_{\theta(g)}(X_0,\varepsilon) = \delta_{\theta(g)}(X_0) + \varepsilon \times \delta^*(X_0) + o(\varepsilon), \tag{22}$$

as $\varepsilon \to 0$; here

$$\delta^*(X_0) = \begin{cases} \sin(\zeta_q), & \text{if } \delta_{\theta(g)}(X_0) = 0, \\ \dfrac{1}{\delta_{\theta(g)}(X_0)} \left(\sum\limits_{j=1}^{q^*} A_j^2 \times \sin^2(\zeta_{k^*+j}) \right)^{1/2}, & \text{if } \delta_{\theta(g)}(X_0) \neq 0, \end{cases} \tag{23}$$

hereinafter, the $o(\cdot)$ symbol in the vector case is understood componentwise.

Proof. Let $X = f(b) \in U_\varepsilon(X_0)$ be an arbitrary point; here $b \in \boldsymbol{B}$. Denote $y = h_g(X)$, then $g(y) = r_{\theta(g)}(X)$ and the Taylor formula yields

$$g(y) = r_{\theta(g)}(X_0) + J_g(y_0) \times (y - y_0) + O(\|X - X_0\|^2). \tag{24}$$

From the definition (15), the value y minimizes the quadtatic form

$$\|X - r_{\theta(g)}(X_0) - J_g(y_0) \times (y - y_0) + O(\|X - X_0\|^2)\|^2,$$

thus, y is the Least Squares solution of the minimization problem (15) and can be written in the form

$$y = y_0 + ((J_g(y_0))^T \times J_g(y_0))^{-1} \times (J_g(y_0))^T \times (X - r_{\theta(g)}(X_0)) + O(\|X - X_0\|^2), \tag{25}$$

and the vector $(X - r_{\theta(g)}(X))$ is orthogonal to the linear space $L_g(y)$:

$$X - r_{\theta(g)}(X) \in (L_g(y))^\perp. \tag{26}$$

Let

$$J_g(y) = Q_g(y) \times \mathrm{Diag}_g(y) \times (V_g(y))^T \tag{27}$$

be the Singular Value Decomposition (SVD) of the $p \times q$ matrix $J_g(y)$; here $Q_g(y)$ is a $p \times q$ orthogonal matrix, $\mathrm{Diag}_g(y)$ and $V_g(y)$ are nondegenerate $q \times q$ diagonal and orthogonal matrices, respectively. Substituing (25) and (27) in (24), we get

$$r_{\theta(g)}(X) = r_{\theta(g)}(X_0) + \pi(y_0) \times (X - r_{\theta(g)}(X_0)) + O(\|X - X_0\|^2),$$

where $\pi(y_0) = Q_g(y_0) \times (Q_g(y_0))^T$ is a projector onto the linear space $L_g(y_0)$ (16).

Taking into account the relation (26) at the point X_0, we obtain

$$\pi(y_0) \times (X - r_{\theta(g)}(X_0)) = \pi(y_0) \times (X - X_0) + \pi(y_0) \times (X_0 - r_{\theta(g)}(X_0)) = \pi(y_0) \times (X - X_0),$$

hence, (24) takes the form

$$r_{\theta(g)}(X = r_{\theta(g)}(X_0) + \pi(y_0) \times (X - X_0) + O(\|X - X_0\|^2),$$

whence comes the relation

$$\begin{aligned} X - r_{\theta(g)}(X) &= (X_0 - r_{\theta(g)}(X_0)) + (X - X_0) - \pi(y_0) \times (X - X_0) + O(\|X - X_0\|^2) \\ &= \Delta(X_0) + \pi^\perp(y_0) \times (X - X_0) + O(\|X - X_0\|^2), \end{aligned} \tag{28}$$

where $\pi^\perp(y_0) = I - \pi(y_0)$ is a projector onto the $(p - q)$-dimensional linear space $(L_g(y_0))^\perp$.

Let

$$J_f(b) = Q_f(b) \times Diag_f(b) \times (V_f(b))^T \tag{29}$$

be the SVD-decomposition of the $p \times q$ matrix $J_f(b)$, where $Q_f(b)$ is a $p \times q$ orthogonal matrix.

Consider the $q \times q$ matrix $(Q_g(y_0))^T \times Q_f(b_0)$, and write its SVD-decomposition in the form

$$(Q_g(y_0))^T \times Q_f(b_0) = O_1 \times Diag(\cos(\zeta)) \times (O_2)^T,$$

where O_1 and O_2 are $q \times q$ orthogonal matrices. The diagonal entries of the diagonal matrix $\mathrm{Diag}(\cos(\zeta))$ are cosines $\cos(\zeta_q), \cos(\zeta_{q-1}), \ldots, \cos(\zeta_1)$ of principal angles between the subspaces $L(X_0)$ and $L_g(y_0)$ arranged in ascending order (17), and the columns of the $p \times q$ orthogonal matrices

$$T_f = Q_f(b_0) \times O_2 \tag{30}$$

and $Q_g(y_0) \times O_1$ are orthonormal bases of these subspaces consisting of the principal vectors $\{t_{f,1}, t_{f,2}, \ldots, t_{f,q}\}$ and $\{t_{g,1}, t_{g,2}, \ldots, t_{g,q}\}$, respectively. Denote

$$\mu(X) = (O_2)^T \times \mathrm{Diag}_f(b_0) \times (V_f(b_0))^T \times (b - b_0) \equiv (\mu_1, \mu_2, \ldots, \mu_q)^T \in R^q,$$

$$\beta(X) = T_f \times \mu(X) \in R^p. \tag{31}$$

From the Taylor series expansion

$$X = X_0 + J_f(b_0) \times (b - b_0) + O(\|X - X_0\|^2),$$

taking into account (29) and the notations (30), (31), we obtain

$$X - X_0 = \beta(X) + O(\|X - X_0\|^2), \tag{32}$$

whence comes that the vectors defined in (31) satisfy the relations

$$\|\beta(X)\|^2 = \|\mu(X)\|^2 = \|X - X_0\|^2 + O(\|X - X_0\|^3).$$

Denote

$$\alpha_0^2(X) = \sum_{j=1}^{k^*} \mu_j^2(X), \quad \alpha_j(X) = \mu_{k^*+j}, \quad j = 1, 2, \ldots, q^*;$$

thus,

$$\alpha_0^2(X) + \sum_{j=1}^{k^*} \alpha_j^2(X) = \|X - X_0\|^2 + O(\|X - X_0\|^3). \tag{33}$$

It also follows from (20) and (32) that

$$\pi^\perp(y_0) \times \beta(X) = \sum_{j=1}^{q^*} \alpha_j(X) \times \sin(\zeta_{k^*+j}) \times e_j,$$

thus

$$\pi^\perp(y_0) \times (X - X_0) = \sum_{j=1}^{q^*} \alpha_j(X) \times \sin(\zeta_{k^*+j}) \times e_j + O(\|X - X_0\|^2),$$

hence, we get from the notation (21) and relation (28) that

$$\Delta(X) = X - r_{\theta(g)}(X) = \Delta^*(X) + O(\|X - X_0\|^2);$$

here

$$\Delta^*(X) = A^* + \sum_{j=1}^{q^*} (A_j + \alpha_j(X) \times \sin(\zeta_{k^*+j})) \times e_j$$

is the main term of the residual vector $\Delta(X)$, and the vector

$$A^* = \Delta(X_0) - \pi_{L^*}(\Delta(X_0)) = \Delta(X_0) - A$$

is an orthogonal $(p - q)$-dimensional complement to the vector A (21). Hence,

$$\delta_{\theta(g)}(X) = \|\Delta(X)\| = \|\Delta^*(X) + O(\|X - X_0\|^2)\|,$$

and

$$\delta_{\theta(g)}(X_0, \varepsilon) = \max\{\|\Delta^*(X) + O(\|X - X_0\|^2)\| : \|X - X_0\| \leqslant \varepsilon\}.$$

Taking into account (33), we write this relation as

$$\delta_{\theta(g)}(X_0,\varepsilon) = \max\left\{\left\|A^* + \sum_{j=1}^{q^*}(A_j + \varepsilon \times \alpha_j \times \sin(\zeta_{k^*+j})) \times e_j + O\left(\varepsilon^2\right)\right\| : \right.$$
$$\left. \alpha_0^2 + \sum_{j=1}^{q^*}\alpha_j^2 \leqslant 1 + O\left(\varepsilon\right)\right\}, \tag{34}$$

here we use normalized variables $\alpha_0 = \alpha_0(X)/\varepsilon$ and $\alpha_j = \alpha_j(X)/\varepsilon, j = 1, 2, \ldots, q^*$.

To find (34), consider first the problem of maximizing the 'main term' in (34), which consists in maximizing the quadratic form

$$\|\Delta^*(X)\|^2 = \|A^*\|^2 + \sum_{j=1}^{q^*}(A_j + \varepsilon \times \alpha_j \times \sin(\zeta_{k^*+j}))^2 \tag{35}$$

under the condition

$$\alpha_0^2 + \sum_{j=1}^{q^*}\alpha_j^2 \leqslant 1,$$

here we used the orthogonal property $A^* \perp L^*$.

We may take the values $\alpha_1, \alpha_2, \ldots \alpha_{q^*}$ to meet the conditions $\mathrm{sgn}(\alpha_j) = \mathrm{sgn}(A_j)$, $j = 1, 2, \ldots, q^*$, then the quantity (35) will increase. Hence, the maximum value of

$$D^2(\alpha_1, \alpha_2, \ldots \alpha_{q^*}) = \sum_{j=1}^{q^*}(A_j + \varepsilon \times \alpha_j \times \sin(\zeta_{k^*+j}))^2 \tag{36}$$

will be reached if $\alpha_0 = 0$ and α_1, $\alpha_2, \ldots \alpha_{q^*}$ satisfy the condition

$$\sum_{j=1}^{q^*}\alpha_j^2 = 1, \tag{37}$$

and we will maximize the quadratic form (36) over $\alpha_1, \alpha_2, \ldots \alpha_{q^*}$ under the condition (37).

Consider separately two cases for the vector A (21):

$$\text{Case 1} : A = 0,$$
$$\text{Case 2} : A \neq 0;$$

the Case 1 includes also the case $\delta_{\theta(g)}(X_0) = 0$.

In Case 1, the optimized function (36) takes the form

$$D^2(\alpha_1, \alpha_2, \ldots \alpha_{q^*}) = \varepsilon^2 \times \sum_{j=1}^{q^*}\alpha_j^2 \times \sin^2(\zeta_{k^*+j}),$$

whose maximum is

$$\max D^2(\alpha_1, \alpha_2, \ldots \alpha_{q^*}) = \varepsilon^2 \times \sin^2(\zeta_q),$$

whence comes that

$$\max\|\Delta^*(X)\|^2 = \delta_{\theta(g)}^2(X_0) + \varepsilon^2 \times \sin^2(\zeta_q) \tag{38}$$

in Case 1; here we use the relation $A^* = \Delta(X_0)$ in the Case 1.

In Case 2, consider the Lagrange function

$$D^2(\alpha_1, \alpha_2, \ldots, \alpha_{q^*}; \lambda) = \sum_{j=1}^{q^*}(A_j + \varepsilon \times \alpha_j \times \sin(\zeta_{k^*+j}))^2 + \lambda \times \left(1 - \sum_{j=1}^{q^*}\alpha_j^2\right),$$

whose optimization gives the following values for variables $\alpha_1, \alpha_2, \ldots, \alpha_{q^*}$:

$$\alpha_j = \varepsilon \times \frac{A_j \times \sin(\zeta_{k^*+j})}{\lambda - \varepsilon^2 \times \sin^2(\zeta_{k^*+j})}, \quad j = 1, 2, \ldots, q^*, \tag{39}$$

where λ is a solution of the equation

$$\varepsilon^2 \times \sum_{j=1}^{q^*} \frac{A_j^2 \times \sin^2(\zeta_{k^*+j})}{\left(\lambda - \varepsilon^2 \times \sin^2(\zeta_{k^*+j})\right)^2} = 1.$$

Denote $x = \varepsilon/\lambda$; then this equation takes the form

$$x^2 \times \sum_{j=1}^{q^*} \frac{A_j^2 \times \sin^2(\zeta_{k^*+j})}{\left(1 - \varepsilon \times x \times \sin^2(\zeta_{k^*+j})\right)^2} = 1,$$

whose solution has the following expansion for small ε:

$$\begin{aligned} x = x_0 + \varepsilon \times x_1 + o(\varepsilon) \equiv & \left(\sum_{j=1}^{q^*} A_j^2 \times \sin^2(\zeta_{k^*+j})\right)^{-1/2} \\ & - \varepsilon \times \frac{\sum_{j=1}^{q^*} A_j^2 \times \sin^4(\zeta_{k^*+j})}{\left(\sum_{j=1}^{q^*} A_j^2 \times \sin^2(\zeta_{k^*+j})\right)^2} + o(\varepsilon), \end{aligned} \tag{40}$$

which correspondes to the maximum value in (36).

Substituting the values (39) and (40) into (36), we obtain

$$\begin{aligned} \max D^2(\alpha_1, \alpha_2, \ldots \alpha_{q^*}) &= \sum_{j=1}^{q^*} \frac{A_j^2}{\left(1 - \varepsilon \times x \times \sin^2(\zeta_{k^*+j})\right)^2} \\ &= \|A\|^2 + 2\varepsilon \times \left(\sum_{j=1}^{q^*} A_j^2 \times \sin^2(\zeta_{k^*+j})\right)^{1/2} \\ &\quad + \varepsilon^2 \times \frac{\sum_{j=1}^{q^*} A_j^2 \times \sin^4(\zeta_{k^*+j})}{\sum_{j=1}^{q^*} A_j^2 \times \sin^2(\zeta_{k^*+j})} + o(\varepsilon^2), \end{aligned}$$

whence comes

$$\begin{aligned} \max \|\Delta^*(X)\|^2 = \delta^2_{\theta(g)}(X_0) &+ 2\varepsilon \times \left(\sum_{j=1}^{q^*} A_j^2 \times \sin^2(\zeta_{k^*+j})\right)^{1/2} \\ &+ \varepsilon^2 \times \frac{\sum_{j=1}^{q^*} A_j^2 \times \sin^4(\zeta_{k^*+j})}{\sum_{j=1}^{q^*} A_j^2 \times \sin^2(\zeta_{k^*+j})} + o(\varepsilon^2) \end{aligned} \tag{41}$$

in Case 2; here we used the equality

$$\delta^2_{\theta(g)}(X_0) = \|A^*\|^2 + \|A\|^2.$$

It follows from (34) that

$$\delta_{\theta(g)}(X_0,\varepsilon) = \max\{\|\Delta^*(X)\| : \sum_{j=1}^{q^*}\alpha_j^2 \leqslant 1\} + o(\varepsilon). \tag{42}$$

Consider separately two cases for the Reconstruction error $\delta_{\theta(g)}(X_0)$ at the point X_0:

$$\text{Case } 1^* : \delta_{\theta(g)}(X_0) = 0,$$
$$\text{Case } 2^* : \delta_{\theta(g)}(X_0) \neq 0;$$

note that $A = 0$ in Case 1^*.

It follows from (38) that in Case 1^*

$$\max\|\Delta^*(X)\| = \varepsilon \times \sin(\zeta_q). \tag{43}$$

Taking into account (41), write the asymptotic expansion of $\max\|\Delta^*(X)\|$ in Case 2^*:

$$\max\|\Delta^*(X)\| = \delta_{\theta(g)}(X_0) + \varepsilon \times \frac{1}{\delta_{\theta(g)}(X_0)}\left(\sum_{j=1}^{q^*} A_j^2 \times \sin^2(\zeta_{k^*+j})\right)^{1/2} + O(\varepsilon^2). \tag{44}$$

Combining relations (42)–(44), we get the relations (22) and (23), which prove Theorem 1. □

2.3 Corollaries of the theorem

From the proof of Theorem 1, one can derive some corollaries. The q-dimensional tangent spaces $L(X_0)$ and $L_g(y_0)$ (16) will be treated as elements of the Grassmann manifold[70] Grass(p,q) composed of q-dimensional linear subspaces in R^p. The maximum principal angle ζ_q between these subspaces defines the metric

$$d_{P,2}(L(X_0), L_g(y_0)) = \sin(\zeta_q) = \|P_L - P_{L^*}\|_2$$

on Grassmann manifold called the projection metric in 2-norm[66], or simply the projection 2-norm[22,28]; here P_L and P_{L^*} are projectors onto the linear spaces $L(X_0)$ and $L_g(y_0)$. In Statistics, this metric is called the Min Correlation Metric[25,33].

If $\zeta_q > 0$, denote by $\eta_{\min} = \sin(\zeta_1)/\sin(\zeta_q) \leqslant 1$ and $\eta_{\min+} = \sin(\zeta_{k^*+1})/\sin(\zeta_q) \leqslant 1$ the sinus of the minimal principal angle ζ_1 and sinus of the minimal positive principal angle ζ_{k^*+1} normalized to the sinus of the maximal principal angle ζ_q, respectively; otherwise put $\eta_{\min} = \eta_{\min+} = 0$.

Obvious inequalities $\|A\| \leqslant \delta_{\theta(g)}(X_0)$ and

$$\|A\| \times \sin(\zeta_{k^*+1}) \leqslant \left(\sum_{j=1}^{q^*} A_j^2 \times \sin^2(\zeta_{k^*+j})\right)^{1/2} \leqslant \|A\| \times \sin(\zeta_q), \tag{45}$$

imply the following Corollary 1.

Corollary 1 of Theorem 1. For $X_0 \in \boldsymbol{X}$ under $\varepsilon \to 0$, we have the following asymptotic upper and lower bounds for the local maximum reconstruction error:

$$\begin{aligned} &\delta_{\theta(g)}(X_0,\varepsilon) \leqslant \delta_{\theta(g)}(X_0) + \varepsilon \times \gamma \times d_{P,2}(L(X_0), L_g(y_0)) + o(\varepsilon), \\ &\delta_{\theta(g)}(X_0,\varepsilon) \geqslant \delta_{\theta(g)}(X_0) + \varepsilon \times \eta_{\min+} \times \gamma \times d_{P,2}(L(X_0), L_g(y_0)) + o(\varepsilon), \end{aligned} \tag{46}$$

where γ is cosine of the angle between the residual vector $\Delta(X_0) = X_0 - r_{\theta(g)}(X_0)$ and its projection onto the linear space L^*; if $\Delta(X_0) = 0$, we put $\gamma = 1$.

From (45), (46), we get a rougher upper bound

$$\delta_{\theta(g)}(X_0,\varepsilon) \leqslant \delta_{\theta(g)}(X_0) + \varepsilon \times d_{P,2}(L(X_0), L_g(y_0)) + o(\varepsilon).$$

From the relations (28), we get Corollary 2, which indicates to what extent the mapping $r_{\theta(g)}$ preserves the local structure of the Data manifold $\boldsymbol{X}$ and characterizes the local non-isometricity of this mapping.

Corollary 2 of Theorem 1. For $X \in U_\varepsilon(X_0)$ and $\varepsilon \to 0$, we have the following asymptotic inequalities:

$$\|(X-X_0)-(r_{\theta(g)}(X)-r_{\theta(g)}(X_0))\| \leqslant \|X-X_0\| \times d_{P,2}(L(X_0), L_g(y_0)) + o(\|X-X_0\|);$$

$$\|(X-X_0)-(r_{\theta(g)}(X)-r_{\theta(g)}(X_0))\| \geqslant \|X-X_0\| \times \eta_{\min} \times d_{P,2}(L(X_0), L_g(y_0)) + o(\|X-X_0\|);$$

$$\|r_{\theta(g)}(X)-r_{\theta(g)}(X_0)\| \leqslant \|X-X_0\| \times \sqrt{1-\eta^2_{\min} \times d^2_{P,2}(L(X_0), L_g(y_0))} + o(\|X-X_0\|);$$

$$\|r_{\theta(g)}(X)-r_{\theta(g)}(X_0)\| \geqslant \|X-X_0\| \times \sqrt{1-d^2_{P,2}(L(X_0), L_g(y_0))} + o(\|X-X_0\|).$$

It follows from Theorem 1 and its corollaries that the greater the distance between the tangent spaces $L(X_0)$ and $L_g(y_0)$, the lower the local generalization ability of the solution θ becomes, the poorer the local structure is preserved, and the poorer the local isometricity properties are ensured at the point $X_0 \in \boldsymbol{X}$.

3 Tangent Bundle Manifold Learning

Denote by

$$L_\theta(r_\theta(X)) = L_g(h(X))$$

the tangent space to the RM $\boldsymbol{X}_\theta$ at the point $r_\theta(X) \in \boldsymbol{X}_\theta$. Thus, it is natural to require in the ML that the procedure θ ensures not only proximity (8) between all the points $X \in \boldsymbol{X}$ and their images $r_\theta(X) \in \boldsymbol{X}_\theta$ but also proximity

$$L(X) \approx L_\theta(r_\theta(X)) \tag{47}$$

between the tangent spaces $L(X) \in \mathrm{Grass}(p,q)$ and $L_\theta(r_\theta(X)) \in \mathrm{Grass}(p,q)$ in some selected metric on the Grassmann manifold $\mathrm{Grass}(p,q)$. The approximate equalities (47) may be treated as **Tangent proximity** between the manifolds $\boldsymbol{X}$ and $\boldsymbol{X}_\theta$.

Requirement of the Tangent proximity in the ML arises in various applications in which ML solution is an intermediate step for some Intelligent Data Analysis problem.

Example 3. Assume that we have to optimize some functional $v(X)$ depending on a p-dimensional vector X lying in a q-dimensional DM $\boldsymbol{X}$ (4) in R^p, $q < p$, covered by a single chart f. By definition, this problems is equivalent to the optimization problem for the functional $v(b) \equiv v(f(b))$ defined on a q-dimensional domain (Coordinate space) $\boldsymbol{B}$.

In applications, an analytical description of the DM $\boldsymbol{X}$ may be unknown, and only a sample $\boldsymbol{X}_n$ from $\boldsymbol{X}$ is available (see Example 1 concerning Wing shape optimization). Based on some solution $\theta = (h, g)$ for the ML, the DM $\boldsymbol{X}$ can be approximated by the RM $\boldsymbol{X}_\theta$ (11) providing the approximate equalities (8) and (13). If the solution θ ensures also the additional 'functional' proximity condition

$$v(r_\theta(X)) \approx v(X), \tag{48}$$

the initial optimization problem for $v(X)$ can be replaced by an optimization problem for the sample-based functional $v_\theta(X) = v(r_\theta(X))$. An amplification of the ML with the functional proximity requirements like (48), called the **Functional DR** (FDR), was stated in Ref. [40]; a neural network based solution for the FDR was proposed in Ref. [7].

The sample-based functional $v_\theta(X)$ can be written in the form

$$v_\theta(X) = v_\theta(g(y)) \equiv v_\theta(y), \quad y = h(X);$$

hence, the initial optimization problem in R^p amounts to the reduced optimization problem for the functional $v_\theta(y)$ depending on the q-dimensional variable $y \in \boldsymbol{Y}_\theta \subset R^q$ belonging to the Reconstructed Coordinate Space $\boldsymbol{Y}_\theta$ (10).

To ensure closeness between specific iterative optimization processes induced by the same optimization gradient-based method applied to the functionals $v(b)$ and $v_\theta(y)$, respectively, it is required to guarantee accurate reconstruction of both the design space $\boldsymbol{X}$ (8), (13) (with FDR requirement (48)) and its tangent spaces $\{L(X) : X \in \boldsymbol{X}\}$ (47). □

In the Manifold theory[42,45], the set

$$TB(\boldsymbol{X}) = \{(X, L(X)) \in \boldsymbol{X} \times \text{Grass}(p,q) : X \in \boldsymbol{X}\},$$

which is composed of points of the manifold $\boldsymbol{X}$ equipped by tangent spaces at these points, is known as the **Tangent Bundle** of the manifold $\boldsymbol{X}$. Thus, accurate reconstruction of the DM $\boldsymbol{X}$ from the sample, which ensures accurate reconstruction of its tangent spaces too, can be considered as reconstruction of the Tangent Bundle TB($\boldsymbol{X}$).

By the above reasons, we propose an amplification of the ML, consisting in accurate reconstruction of the tangent bundle TB($\boldsymbol{X}$) from the sample $\boldsymbol{X}_n$, which will be referred to as the **Tangent Bundle Manifold Learning** (TBML) problem.

A strict definition of the TBML is as follows: Given an input dataset $\boldsymbol{X}_n$ (1) sampled from a q-dimensional DM $\boldsymbol{X}$ (4) covered by a single chart, construct a **TBML-solution** $\theta = (h, g)$ consisting of an Embedding mapping h (3) and a Reconstruction mapping g (5) which determines the **Reconstructed Tangent Bundle**

$$TB_\theta(\boldsymbol{X}_\theta) = \{(X', L_\theta(X')) : X' \in \boldsymbol{X}_\theta\} \equiv \{(g(y), L_g(y)) : y \in \boldsymbol{Y}_\theta = h(\boldsymbol{X})\}$$

of the manifold $\boldsymbol{X}_\theta$ (11) and ensures its proximity to the Tangent bundle TB($\boldsymbol{X}$):

$$TB(\boldsymbol{X}) \approx TB_\theta(\boldsymbol{X}_\theta), \tag{49}$$

which means both the Manifold proximity $\boldsymbol{X} \approx \boldsymbol{X}_\theta$ (13) and the Tangent proximity

$$\boldsymbol{L} = \{L(X), X \in \boldsymbol{X}\} \approx \boldsymbol{L}_\theta = \{L_\theta(X'), X' \in \boldsymbol{X}_\theta\} \equiv \{L_g(y) : y \in \boldsymbol{Y}_\theta = h(\boldsymbol{X})\}; \tag{50}$$

the latter proximity means proximity (47) for all the points $X \in \boldsymbol{X}$ in some selected metric on the Grassmann manifold Grass(p,q).

Note, that the sets $\boldsymbol{L}$ and $\boldsymbol{L}_\theta$ are q-dimensional submanifolds in the Grassmann manifold Grass(p,q); we will call them the **Tangent Manifold** and **Reconstructed Tangent Manifold**, respectively. The Tangent proximity (50) means that the linear spaces $L_\theta(r_\theta(X)) = L_g(h(X)) \in \boldsymbol{L}_\theta$ accurately reconstruct the linear spaces $L(X) \in \boldsymbol{L}$.

From the Statistical point of view, the defined TBML may also be considered as the following Estimation Problem: from a finite dataset $\boldsymbol{X}_n$ randomly sampled from smooth q-dimensional manifold $\boldsymbol{X}$ with Tangent manifold $\boldsymbol{L}$, estimate $\boldsymbol{X}$ and $\boldsymbol{L}$.

Note. The term ‘tangent bundle’ is used in the ML for various purposes: as the approximation of manifold shape from the data in Ref. [23]; for geometric interpretation of the Contractive Auto-Encoder[50] algorithm in Ref. [51]; as the name of the ‘Tangent Bundle Approximation’ algorithm for the Tangent Space Learning Problem in Ref. [57,58,59], etc. The above given TBML definion is new and different from all known use of this term in the ML.

4 Grassman&Stiefel Eigenmaps Algorithm

This section describes the proposed solution for the TBML, called Grassman&Stiefel Eigenmaps (GSE), presented by the below described GSE-algorithm.

4.1 Preliminaries

Hereinafter, by numbers $\{\varepsilon_i > 0, i = 1, 2, \ldots\}$ we denote the algorithm parameters. The GSE, as well as other 'local' algorithms[15, 47], includes the following standard auxiliary stages, in which:

- For arbitrary point $X \in \boldsymbol{X}$, ε_1-ball $U_E(X)$ centered at X is constructed as follows:

  $U_E(X) = \{X' \in \boldsymbol{X}_n : \|X' - X\| < \varepsilon_1\}$ for sample points $X \in \boldsymbol{X}_n$; if X is OoS point ($X \notin \boldsymbol{X}_n$), X is included in the $U_E(X)$ also.

  Following Ref. [2], define the Euclidean 'heat' kernel
  $$K_E(X, X') = K_0(X, X') \times \exp\{-\|\varepsilon_2 \times (X - X')\|^2\}, \tag{51}$$
  where $K_0(X, X') = 1$ if $X \in U_E(X')$ and $X' \in U_E(X)$, and $K_0(X, X') = 0$ otherwise.

- The PCA-based approximations $L_{PCA}(X)$ for the tangent spaces $L(X)$ are constructed as follows. By applying the Weighted PCA[39] with the weights (51) to the set $U_E(X)$, the ordered eigenvalues $\lambda_1(X) \geqslant \lambda_2(X) \geqslant \ldots \geqslant \lambda_p(X)$ and the corresponding p-dimensional orthonormal principal vectors are constructed.

Define the orthogonal $p \times q$ matrix $Q_{PCA}(X)$ with the columns consisting of the first q principal vectors, and denote
$$L_{PCA}(X) = \mathrm{Span}(Q_{PCA}(X)) \in \mathrm{Grass}(p, q) \tag{52}$$
the linear space spanned by columns of the $Q_{PCA}(X)$; this linear space can be considered as the PCA-approximation of the tangent space $L(X)$.

Note. In Ref. [29,65] and others works, various methods for a choice of the local neighborhoods for applying the PCA are proposed.

Denote
$$\boldsymbol{X}_h = \{X \in R^p : \lambda_q(X) > \varepsilon_3\};$$
this set will be the domain of definition of the built in the future Embedding mapping h (3).

In what follows, we assume that the DM $\boldsymbol{X}$ is well sampled to provide the inclusion $\boldsymbol{X} \subset \boldsymbol{X}_h$. If $X \in \boldsymbol{X}_h$ and the neighborhood $U_E(X)$ is small enough, then[1]
$$L(X) \approx L_{PCA}(X). \tag{53}$$

In the paper, we will consider the aggregate kernel
$$K(X, X') = K_E(X, X') \times K_G(X, X'), \tag{54}$$
in which
$$K_G(X, X') = K_{BC}(L_{PCA}(X), L_{PCA}(X')) = Det^2[(Q_{PCA}(X))^T \times Q_{PCA}(X')]; \tag{55}$$
here $K_{BC}(\cdot,\cdot)$ is the Binet-Cauchy kernel[28] on the Grassmann manifold $\mathrm{Grass}(p, q)$ induced by the Binet-Cauchy metric[28,69]
$$d_{BC}(L_{PCA}(X), L_{PCA}(X')) = \{1 - Det^2[(Q_{PCA}(X))^T \times Q_{PCA}(X')]\}^{1/2} \tag{56}$$

on the Grass(p, q).

Note. Among all known metrics on the Grassmann manifold Grass(p, q), are only two metrics between the linear spaces $L, L' \in$ Grass(p, q) that induce corresponding kernels on the Grassmann manifold: the Binet-Cauchy metric d_{BC} (56) and the Projection metric[28] $d_{P,F}(L, L') = 2^{-1/2} \times \|P_L - P_{L'}\|_F$, also called the Projection F-norm[22] and Chordal distance[17]. There are other reasons also to use the Binet-Cauchy metric (56) and kernel $K_G(X, X')$ (55) in the GSE-algorithm.

The aggregate kernel (54) reflects not only geometrical nearness between X and X' but also nearness between the linear spaces $L_{PCA}(X)$ and $L_{PCA}(X')$, whence comes, when (53), a nearness between the tangent spaces $L(X)$ and $L(X')$.

The rest of this section is organized as follows. First, we describe a structure and give an justification of the GSE (Subsection 4.2). The next subsections 4.3–4.5 contain the details of the GSE algorithm. Subsection 4.6 presents some properties of this algorithm.

4.2 Structure and justification of the GSE

The GSE consists of two successively performed main parts:

- **Approximation of the Tangent manifold**. A sample-based submanifold
$$\boldsymbol{L}_H = \{L_H(X), X \in \boldsymbol{X}\} \subset \mathrm{Grass}(p, q)$$
of the Grassmann manifold Grass(p, q) consisting of the linear spaces $L_H(X)$ is constructed to approximate the Tangent manifold $\boldsymbol{L}$. The submanifold $\boldsymbol{L}_H$ is defined by the family $\boldsymbol{H} = \{H(X), X \in \boldsymbol{X}\}$ consisting of the $p \times q$ matrices $H(X)$ smoothly depending on $X \in \boldsymbol{X}$: the linear space $L_H(X) = \mathrm{Span}(H(X))$ is spanned by columns of the matrix $H(X)$.

- **Reconstructing the Tangent Bundle of the Data Manifold**. Given the constructed submanifold $\boldsymbol{L}_H$, TBML-solution θ is constructed in such a way to ensure the following properties: the RM $\boldsymbol{X}_\theta$ (11) approximates the DM $\boldsymbol{X}$ and the Reconstructed Tangent Manifold $\boldsymbol{L}_\theta$ close to the given submanifold $\boldsymbol{L}_H$ (which, in turn, close by construction to the Tangent manifold $\boldsymbol{L}$), whence comes the Tangent bundle proximity.

Describe briefly the idea and justification of the proposed GSE-algorithm.

Approximation of the Tangent manifold. Given the constructed PCA-based linear spaces $L_{PCA}(X)$ (52), the matrices $H(X)$ are constructed to meet the relations
$$L_H(X) = L_{PCA}(X)$$
for all $X \in \boldsymbol{X}_h$; whence, when (53), comes the required proximity
$$L_H(X) \approx L(X). \tag{57}$$

Let Stief(p, q) denotes non-compact Stiefel manifold[22,45], consisting of all tall-skinny $p \times q$ matrices M with $\mathrm{Rank}(M) = q, q \leqslant p$. To achieve a smootheness of the mapping
$$H : X \in \boldsymbol{X}_h \to H(X) \in \mathrm{Stief}(p, q)$$
defined on the domain $\boldsymbol{X}_h \subset R^p$, the preliminary matrix set $\boldsymbol{H}_n \subset$ Stief(p, q) consisting of the matrices $H_i \in$ Stief(p, q) that meet the constraints
$$\mathrm{Span}(H_i) = L_{PCA}(X_i), \quad i = 1, 2, \ldots, n, \tag{58}$$

is chosen to minimize the quadratic form

$$\Delta_n(\boldsymbol{H}_n) = \frac{1}{2}\sum_{i,j=1}^{n} K(X_i, X_j) \times \|H_i - H_j\|_F^2 \tag{59}$$

under the normalizing constraint

$$\frac{1}{K}\sum\nolimits_{i=1}^{n} K(X_i) \times (H_i^T \times H_i) = I_q; \tag{60}$$

here I_q is $q \times q$ identity matrix, and

$$K(X) = \sum\nolimits_{j=1}^{n} K(X, X_j), \quad i = 1, 2, \ldots, n; \quad K = \sum\nolimits_{i=1}^{n} K(X_i).$$

Then, based on the constructed matrix set $\boldsymbol{H}_n$, the value $H(X) \in \mathrm{Stief}(p, q)$ for arbitrary point X that meets the constraint

$$\mathrm{Span}(H(X)) = L_{PCA}(X), \tag{61}$$

is constructed by minimizing the quadratic form

$$\Delta_H(H(X)) = \sum\nolimits_{j=1}^{n} K(X, X_j) \times \|H(X) - H_j\|_F^2. \tag{62}$$

The solutions of the optimization problems (59) and (62) are described in the Subsection 4.3.

Reconstructing the Tangent Bundle of the Data Manifold. The desired mappings h (3) and g (5) will build in the future in such a way that Jacobian $J_g(h(X))$ of the mapping $g(y)$ at the point $y = h(X)$ is close to the matrix $H(X)$:

$$J_g(h(X)) \approx H(X) \tag{63}$$

for all the points $X \in \boldsymbol{X}$. The Taylor series expansion of the function $g(y)$:

$$g(y') - g(y) \approx J_g(y) \times (y' - y) \tag{64}$$

for near points $y, y' \in \boldsymbol{Y}_g$, under the desired condition (63), gives the following relation

$$g(h(X')) - g(h(X)) \approx H(X) \times (h(X') - h(X)) \tag{65}$$

for near points $X, X' \in \boldsymbol{X}$. Under the desired conditions $g(h(X)) \approx X$ (8), from here comes the relation

$$X' - X \approx H(X) \times (h(X') - h(X)) \tag{66}$$

for near points $X, X' \in \boldsymbol{X}$. Under the already constructed family $\boldsymbol{H}$, these approximate equalities can be considered as regression equations for the embeddings $h(X)$.

At first, consider the regression equations (66) for near sample points $X, X' \in \boldsymbol{X}_n$ and compute the preliminary vector set $\boldsymbol{h}_n = \{h_1, h_2, \ldots, h_n\} \subset R^q$ as standard least squares solution which minimizes the weighted residual

$$\Delta_n(\boldsymbol{h}_n) = \sum\nolimits_{i,j=1}^{n} K(X_i, X_j) \times |(X_j - X_i) - H(X_i) \times (h_j - h_i)|^2 \tag{67}$$

under natural normalizing constraint

$$h_1 + h_2 + \ldots + h_n = 0 \in R^q. \tag{68}$$

Then, under the already constructed vector set $\boldsymbol{h}_n$, choose the value $h(X)$ for arbitrary point $X \in \boldsymbol{X}$ by minimizing the weighted residual

$$\Delta_h(h(X)) = \sum\nolimits_{j=1}^{n} K(X, X_j) \times |(X_j - X) - H(X) \times (h_j - h(X))|^2. \tag{69}$$

The solutions of the optimization problems (67) and (69) will be given in the Subsection 4.4.

In the Subsection 4.5, the nearness measure (kernel) $k(y, y')$ in the Reconstructed Coordinate space $\boldsymbol{Y}_h = h(\boldsymbol{X}_h)$ between the points $y \in \boldsymbol{Y}_h$ and $y' \in \boldsymbol{Y}_n = \{h(X), X \in \boldsymbol{X}_n\}$ is introduced to provide the approximate equalities

$$k(h(X), h(X')) \approx K(X, X'). \tag{70}$$

for the near points $X \in \boldsymbol{X}_h$ and $X' \in \boldsymbol{X}_n$.

The matrices $\{G(y)\}$ are constructed to ensure the approximate equalities

$$G(h(X)) \approx H(X); \tag{71}$$

thus, the relations (64)–(66) can be written in the form

$$X - g(y) \approx G(y) \times (h(X) - y) \tag{72}$$

for near points y and $h(X)$.

By writing these equations for the sample points $X \in \boldsymbol{X}_n$ whose embedding $h(X)$ close to the point y, we can estimate the value $g(y)$ for 'OoS points' $y \in \boldsymbol{Y}_h/\boldsymbol{Y}_n$ to provide both the desired relations (8) and

$$G(y) = J_g(y). \tag{73}$$

Therefore, the relations (57), (63), (71)–(73) provide the required proximities (13) and (50), whence comes the Tangent Bundle proximity (49).

Note 1. In the Laplacian Eigenmaps[2] solution of the Sample Embedding Problem, the embeddings $h_i = h_{(n)}(X_i)(2), i = 1, 2, \ldots, n$, are chosen to minimize the cost function

$$L_{LE}(h_{(n)}) = \sum\nolimits_{i,j=1}^{n} K_E\left(X_i, X_j\right) \times \|h_i - h_j\|^2 \tag{74}$$

under the normalizing constraint

$$\sum\nolimits_{i=1}^{n} K_E\left(X_i\right) \times \left(h_i \times h_i^T\right) = I_q, K_E(X) = \sum\nolimits_{j=1}^{n} K_E\left(X, X_j\right)$$

required to avoid a degenerate solution. The minimized cost function (59) similar to the cost function (74), but minimization in (59) is done over the matrices $\{H_i \in \text{Stief}(p, q)\}$ under the constraints (58) and (60), while minimization in (74) is done over the vectors $\{h_i \in R^q\}$ under the constraint (68). Note also that the aggregate kernel K (54) is used in (59), while the Euclidean heat kernel K_E (51) is used in (74).

The cost function (74), called the Laplacian[2], is discrete analogous to the continuous functional

$$L(h) = \int_{\mathbf{X}} \|\nabla h\left(X\right)\|^2 = \sum\nolimits_{i=1}^{q} \int_{\mathbf{X}} L_{LB}\left(h_i\right) \times h_i \tag{75}$$

defined on the vector embedding $h(X) = (h_1(X), h_2(X), \ldots, h_q(X))^T$ (3), which, in turn, is defined on the Data Manifold $\boldsymbol{X}$; here $L_{LB}(h_i) = -\text{div}(\nabla h_i)$ is the Laplace Beltrami operator on manifold applied to the component h_i of the Embedding function h.

The cost function (59) is also discrete analogous to the continuous functional $L(\text{Hess}(g))$, which is similar to the functional $L(h)$ (75), but in which the Laplace Beltrami operator is applied to the elements of all the Hessian matrices $\text{Hess}_{g,k}(y)$ of the components $g_k(y)$ of the vector function $g(y) = (g_1(y), g_2(y), \ldots, g_p(y))^T$ (5) at the point $y = h(X)$.

Note 2. Under the condition (63), the typical term $\|H\left(X'\right) - H(X)\|_F^2$ in the quadratic form (59) equals approximately to the quantity

$$\sum\nolimits_{k=1}^{p} \left|\text{Hess}_{g,k}\left(h\left(X\right)\right) \times \left(h(X') - h\left(X\right)\right)\right|^2;$$

thus, the quadratic form $\Delta_n(\boldsymbol{H}_n)$ (59) is close to the quadratic form

$$\Delta_{n,\theta}(\boldsymbol{X}_n) = \sum_{i>j} K(X_i, X_j) \times \left\{ (y_i - y_j)^T \times \left[H_\theta^2(y_i) + H_\theta^2(y_j) \right] \times (y_i - y_j) \right\},$$

where $y_i = h(X_i), i = 1, 2, \ldots, n$, and $q \times q$ matrix $H_\theta^2(y)$ is

$$H_\theta^2(y) = \sum\nolimits_{k=1}^{p} (\text{Hess}_{g,k}(y))^T \times \text{Hess}_{g,k}(y),$$

whose trace is squared Frobenius norm of the Hessian of Reconstruction mapping $g(y)$ (5).

Thus, the minimum values of the quadratic forms (59) and (62) characterize the averaged local curvature of the reconstructed manifold $\boldsymbol{X}_\theta$.

Note 3. Tangent Space Learning[29], or, more generally, Subspace Learning[11], is newly emerging direction in the ML. The problem of estimating the tangent spaces $L(X)$ to the DM $\boldsymbol{X}$ at the points $X \in \boldsymbol{X}$ in the form of a smooth function of the point X was considered in some previous works. The matrices whose rows approximately span the tangent spaces were constructed using Artificial Neural Networks with one hidden layer[5] or Radial Basis Functions[19,20]. In Ref. [29], other method (namely, Persistent Tangent Space Learning) is proposed for constructing the approximations for the tangent spaces, which smoothly varied on the manifold; this method also based on considering the tangent spaces as points in Grassmann manifold.

4.3 Sample-based approximation of Tangent Manifold

The Grassmann manifold Grass(p, q) can be considered[22] as the quotient manifold of the Stiefel manifold Stief(p, q) with respect to the group Stief(q, q). Thus, the matrices $H_i \in$ Stief(p, q), $i = 1, 2, \ldots, n$, and $H(X) \in$ Stief(p, q), which satisfy the conditions (58) and (61), respectively, can be presented in the form

$$H(X) = Q_{PCA}(X) \times v(X), \tag{76}$$

$$H_i = Q_{PCA}(X_i) \times v_i, \quad i = 1, 2, \ldots, n, \tag{77}$$

where $v_1, v_2, \ldots, v_n, v(X) \in$ Stief(q, q).

From (77), the quadratic form (59) can be written as

$$\Delta_V(v_1, v_2, \ldots, v_n) = \frac{1}{2} \sum\nolimits_{i,j=1}^{n} K(X_i, X_j) \times \| Q_{PCA}(X_i) \times v_i - Q_{PCA}(X_j) \times v_j \|_F^2. \tag{78}$$

Note. Minimization of (i, j)-term in (78) over the orthogonal matrices v_i and v_j is known as the Procrustes problem[22,27]. Thus, minimization of the quadratic form (78) over the matrices $v_1, v_2, \ldots, v_n$ may be referred to as the Averaged Procrustes problem.

Consider nq×nq matrices $\Phi = \|\Phi_{ij}\|$ and $F = \|F_{ij}\|$ which consist of n^2 $q \times q$ matrices

$$\Phi_{ij} = K(X_i, X_j) \times (Q_{PCA}(X_i))^T \times Q_{PCA}(X_j),$$

$$F_{ij} = \delta_{ij} \times \frac{1}{K} \times K(X_i) \times I_q, \quad i, j = 1, 2, \ldots, n.$$

Using the representation (77), the quadratic form (59) under the constraint (60) can be wtitten in the form

$$\Delta_V(v_1, v_2, \ldots, v_n) = K - Tr(\boldsymbol{V}^T \times \Phi \times \boldsymbol{V}),$$

where the transposed value

$$\boldsymbol{V}^T = \left(v_1^T : v_2^T : \cdots : v_n^T \right) \tag{79}$$

of the $(nq) \times q$ matrix $\boldsymbol{V}$ is the n sequentially recorded transposed submatrices $v_1, v_2, \ldots, v_n$.

The constraint (60), which coincides with the constraint

$$\frac{1}{K} \sum\nolimits_{i=1}^{n} K(X_i) \times (v_i^T \times v_i) = I_q,$$

can be wtitten in the form

$$\boldsymbol{V}^T \times F \times \boldsymbol{V} = I_q. \tag{80}$$

Thus, the optimization problem (59) is reduced to maximizing the quadratic form $\mathrm{Tr}(\boldsymbol{V}^T \times \Phi \times \boldsymbol{V})$ over the $(nq) \times q$ matrices $\boldsymbol{V}$ under the constraint (80), and we obtain the following Theorem 2.

Theorem 2. Let the columns of $(nq) \times q$ matrix $\boldsymbol{V}$ are the orthonormal eigenvectors $V_1, V_2, \ldots, V_q \in R^{nq}$ corresponding to the q largest eigenvalues in the generalized eigenvector problem

$$\Phi \times V = \lambda \times F \times V.$$

Let $v_1, v_2, \ldots, v_n$ be the matrices whose transposed values are defined as the n sequentially recorded submatrices in the representation (79) of the constructed matrix $\boldsymbol{V}^T$.

Then the matrix set $\boldsymbol{H}_n$ consisting of the matrices (77) with the above defined matrices $v_1, v_2, \ldots, v_n$ minimizes the quadratic form (59) under the constraint (60).

Note. The constructed linear spaces $\{\mathrm{Span}(H_i)\}$ can be considered as the result of a global alignment of the PCA-based linear spaces $\{L_{PCA}(X_i)\}$. Similar alignment problem was studied in the LTSA[72,73] with using a cost function which differs from our cost function (59), (78).

Given the already constructed matrix set $\boldsymbol{H}_n$, a solution of the minimization problem (62) for the quadratic form $\Delta_H(H(X))$ is obtained in explicit form in Theorem 3.

Theorem 3. The matrix

$$H(X) \equiv H(X|\boldsymbol{H}_n) = \pi(X) \times H_{KNR}(X) \tag{81}$$

satisfies the constraint (61) and minimizes the quadratic form $\Delta_H(H(X))$ (62); here

$$\pi(X) = Q_{PCA}(X) \times (Q_{PCA}(X))^T \tag{82}$$

is projector onto the linear space $L_{PCA}(X)$ (52), and

$$H_{KNR}(X) = \frac{1}{K(X)} \sum\nolimits_{j=1}^{n} K(X, X_j) \times H_j$$

is standard Kernel Non-parametric Regression[67]-based estimator for $H(X)$ based on the preliminary values $H_j \in \boldsymbol{H}_n$ of the matrix $H(X)$ at the sample points.

Corollary of Theorem 3. The matrix $v(X)$ in the representation (76) of the matrix $H(X)$ equals to

$$v(X) = (Q_{PCA}(X))^T \times H_{KNR}(X). \tag{83}$$

The final values $\{H(X_i)\}$ (81) at the sample points do not coincide with the preliminary values $\{H_i \in \boldsymbol{H}_n\}$. But the followibg Theorem 4 hold true.

Theorem 4. The set $\boldsymbol{H}_n$, which was constructed in the Theorem 2, minimizes the averaged residual

$$D(\boldsymbol{H}_n) = \sum\nolimits_{i=1}^{n} K(X_i) \times \|H_i - H(X_i|\boldsymbol{H}_n)\|_F^2$$

over the matrices $\{H_i\}$ (77) under the normalizing constraint (60).

Note. Such approach to constructing the matrices $\{H_i\}$ by minimizing the residual $D(\boldsymbol{H}_n)$ have parallels with the LLE[52]; an equivalence of the optimization schemes $\Delta_n(\boldsymbol{H}_n)$ and $D(\boldsymbol{H}_n)$ like the equivalence LLE and LE[2] set out in Ref. [2].

4.4 Solution of the Embedding Problem

Under the representation (77), the weighted residual $\Delta_n(\boldsymbol{h}_n)$ (67) can be written as

$$\Delta_n(\boldsymbol{h}_n) = \frac{1}{2}\sum\nolimits_{i,j=1}^{n} K(X_i, X_j) \times \left\{\left|\pi^{\perp}(X_i) \times (X_j - X_i)\right|^2 + \left|v_i \times (h_j - h_i) - c_{j|i}\right|^2\right\},$$

where $c_{j|i} = (Q_{PCA}(X_i))^T \times (X_j - X_i)$ are the expansion coefficients of projection of the vectors $(X_j - X_i)$ onto the linear space $L_{PCA}(X_i)$ in the PCA basis $Q_{PCA}(X_i)$. Thus, we obtain the following Theorem 5.

Theorem 5. The vector set $\boldsymbol{h}_n$, which minimizes the weighted residual $\Delta_n(\boldsymbol{h}_n)$ (67), is the solution of the linear least squares equations with j^{th} equation

$$\sum_{i=1}^{n} K(X_i, X_j) \times (v_i^T \times v_i + v_j^T \times v_j) \times (h_j - h_i) = \sum_{i=1}^{n} K(X_i, X_j) \times \left(v_i^T \times c_{j|i} - v_j^T \times c_{i|j}\right),$$

$j = 1, 2, \ldots, n$, complemented by the normalizing equation (68).

Note. The vector set $\boldsymbol{h}_n$ determined in Theorem 5 gives a new solution of the Sample Embedding Problem.

Given the already constructed vector set $\boldsymbol{h}_n$, the solution of the minimization problem (69) for the weighted residual $\Delta_h(h(X))$ is obtained in explicit form in Theorem 6.

Theorem 6. The vector

$$h(X) = h_{KNR}(X) + v^{-1}(X) \times (Q_{PCA}(X))^T \times \left(X - \frac{1}{K(X)}\sum_{j=1}^{n} K(X, X_j) \times X_j\right) \tag{84}$$

minimizes the weighted residual $\Delta_h(h(X))$ (69); here $v(X)$ is determined in (83) and

$$h_{KNR}(X) = \frac{1}{K(X)}\sum\nolimits_{j=1}^{n} K(X, X_j) \times h_j$$

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

Note. The embedding $h(X)$ determined in Theorem 6 gives a new solution of the Manifold Embedding Problem.

4.5 Tangent Bundle reconstruction

4.5.1 Nearness measure in the Reconstructed Coordinate space

Let the Sample embedding dataset

$$\boldsymbol{Y}_n = h(\boldsymbol{X}_n) = \{h(X_1), h(X_2), \ldots, h(X_n)\} \equiv \{y_1, y_2, \ldots, y_n\}$$

consists of values of the mapping h (84) at the sample points. Denote the inverse mapping $h^{-1}(y)$ defined only on the sample embeddings $y \in \boldsymbol{Y}_n$ as $h^{-1}(y_i) = X_i, i = 1, 2, \ldots, n$.

Let $y = h(X) \in \boldsymbol{Y}_h = h(\boldsymbol{X}_h)$ be an arbitrary point, and let $X' \in \boldsymbol{X}_n$ be some sample point close to X. It is follows from (66) that

$$X - X' \approx H(X') \times (y - y'),$$

where $y' = h(X') \in \boldsymbol{Y}_n$, whence, with taking into account the relations (77), we get the approximate relation

$$|X - X'| \approx |v(X') \times (y - y')|. \tag{85}$$

Denote $q \times q$ matrix $K_v(X) = v^T(X) \times v(X)$, and define

$$u_E(y) = \{y' \in \boldsymbol{Y}_n : [(y' - y)^T \times K_v(h^{-1}(y')) \times (y' - y)]^{1/2} < \varepsilon_1\}$$

the neighborhood of the point $y \in \boldsymbol{Y}_h$. Introduce the Euclidean nearness measure

$$k_E(y, y') = k_0(y, y') \times \exp\{-(\varepsilon_2)^2 \times (y - y')^T \times K_v(h^{-1}(y')) \times (y - y')\} \tag{86}$$

between the points $y \in \boldsymbol{Y}_h$ and $y' \in \boldsymbol{Y}_n$; here $k_0(y, y') = 1$ if $y' \in U_E(y)$, and $k_0(y, y') = 0$ otherwise. From (85), (86), we get the approximate equalities

$$K_E(X, X') \approx k_E(h(X), h(X')) \tag{87}$$

for close points $X \in \boldsymbol{X}_h$ and $X' \in \boldsymbol{X}_n$.

By applying the PCA to the set

$$U_E^*(y) = \{h^{-1}(y') : y' \in u_E(y)\} \subset \boldsymbol{X}_n,$$

the ordered eigenvalues $\lambda_1^*(y) \geqslant \lambda_2^*(y) \geqslant \ldots \lambda_p^*(y)$ and the corresponding principal vectors are constructed. Introduce the subset

$$\boldsymbol{Y}_g = \{y \in R^q : \lambda_q^*(y) > \varepsilon_3\},$$

which will be the domain of definition of the built in the future mapping g and matrix G.

For $y \in \boldsymbol{Y}_g$, define the $p \times q$ orthogonal matrix $q(y)$ with columns consisting of the first q principal vectors, and define the linear space

$$L^*(y) = \mathrm{Span}(q(y)) \in \mathrm{Grass}(p, q).$$

Using the Binet-Cauchy kernel (55) on the Grassmann manifold Grass(p, q), introduce the tangent nearness measure between the points $y \in \boldsymbol{Y}_g$ and $y' \in \boldsymbol{Y}_n$ by the kernel

$$k_G(y, y') = Det^2(q^T(y) \times Q_{PCA}(h^{-1}(y'))), \tag{88}$$

and introduce the aggregate kernel

$$k(y, y') = k_E(y, y') \times k_G(y, y'). \tag{89}$$

on the Reconstructed Coordinate space.

As before, we assume that the DM $\boldsymbol{X}$ is well sampled to provide inclusion $\boldsymbol{Y}_\theta \subset \boldsymbol{Y}_g$. It is follows from (85), (87) and the introduced definitions (88), (89), that the approximate equalities (70) hold for the near points $X \in \boldsymbol{X}_h$ with $h(X) \in \boldsymbol{Y}_g$ and $X' \in \boldsymbol{X}_n$.

4.5.2 Tangent Manifold Reconstruction

Based on the desired conditions (71) and (73), construct $p \times q$ matrix $G(y)$ for arbitrary point $y \in \boldsymbol{Y}_g$, which satisfies the constraint $\mathrm{Span}(G(y)) = L^*(y)$ and minimizes the quadratic form

$$\Delta_G(G(y)) = \frac{1}{2} \sum\nolimits_{j=1}^{n} k\left(y, y_j\right) \times \left\| G\left(y\right) - H\left(X_j\right) \right\|_F^2. \tag{90}$$

A solution of this problem in explicit form is obtained in Theorem 7.

Theorem 7. The matrix

$$G(y) = q(y) \times q^T(y) \times \frac{1}{k\left(y\right)} \sum\nolimits_{j=1}^{n} k\left(y, y_j\right) \times H\left(X_j\right), \quad k(y) = \sum\nolimits_{j=1}^{n} k\left(y, y_j\right),$$

meets the required constraint and minimizes the quadratic form (90).

4.5.3. Data Manifold Reconstruction

For near points $y \in \boldsymbol{Y}_g$ and $y' \in \boldsymbol{Y}_n$, the approximate relations (64) under the desired conditions (8) and (73) can be written in the form

$$h^{-1}(y') - g(y) \approx G(y) \times (y' - y).$$

Construct the vector $g(y) \in R^p$ for an arbitrary point $y \in \boldsymbol{Y}_g$ by minimizing the weighted residual

$$\Delta_g(g(y)) = \sum\nolimits_{j=1}^{n} k\left(y, y_j\right) \times \| X_j - g\left(y\right) - G\left(y\right) \times \left(y_j - y\right) \|^2. \tag{91}$$

A solution of this problem in explicit form is obtained in Theorem 8.

Theorem 8. The p-dimensional vector

$$g(y) = g_{KNR}(y) + G(y) \times \left(y - \frac{1}{k\left(y\right)} \sum\nolimits_{j=1}^{n} k\left(y, y_j\right) \times y_j \right) \tag{92}$$

satisfies the condition $J_g(y) = G(y)$ (73) for the points $y \in \boldsymbol{Y}_g$ and minimizes the weighted residual (91); here

$$g_{KNR}(y) = \frac{1}{k\left(y\right)} \sum\nolimits_{j=1}^{n} k\left(y, y_j\right) \times X_j$$

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

Note. For near points $y \in \boldsymbol{Y}_g$ and $y' \in \boldsymbol{Y}_n$, the approximate relations (64) can be also written in the form

$$g(y) - h^{-1}(y') \approx H(h^{-1}(y')) \times (y - y'),$$

and the weighted residual (91) can be replaced by close quadratic form

$$\Delta(g(y)) = \sum\nolimits_{j=1}^{n} k\left(y, y_j\right) \times \| g\left(y\right) - X_j - H\left(X_j\right) \times \left(y - y_j\right) \|^2,$$

whose optimization gives the solution

$$g^*(y) = g_{KNR}(y) + \frac{1}{k\left(y\right)} \sum\nolimits_{j=1}^{n} k\left(y, y_j\right) \times H(X_j) \times (y - y_j), \tag{93}$$

which is close to the solution $g(y)$ (92).

The mapping $g(y)$ (92) and the close mapping $g^*(y)$ (93) determine the Reconstructed Data Manifolds $\boldsymbol{X}_g$ and $\boldsymbol{X}_{g^*}$ (9), (11).

4.6 Properties of the GSE

We present without proof a few properties of the GSE algorithm.

1) If the sample size n tends to infinity together with a tending the threshold $\varepsilon_1 = \varepsilon_{1n}$ in the neighborhoods $U_E(X)$ and $u_E(y)$ to 0 with an appropriate convergence rate, then all the points from the sets $\boldsymbol{X}$ and $\boldsymbol{Y}_\theta$ will fall into the sets $\boldsymbol{X}_h$ and $\boldsymbol{Y}_g$, respectively.

2) As was showed in Ref. [71], the radius ε_{1n} must tend to 0 as $n \to \infty$ with the rate $O(n^{-1/(q+2)})$; this rate ensures asymptotically optimal Tangent Bundle proximity:

$$\|X - r_\theta(X)\| = O(n^{-2/(q+2)}) \text{ and } d_{P,2}(L(X), L_\theta(r_\theta(X))) = O(n^{-1/(q+2)}),$$

uniformly over all the points $X \in \boldsymbol{X}$; here $d_{P,2}$ is the projection 2-norm metric on the Grassmann manifold Grass(p, q).

3) The residual vector $\Delta(X) = X - r_\theta(X)$ can be approximately represented in the form

$$\Delta(X) \approx \pi^{\perp}(X) \times (X - \tau(X))$$

with the error of the order $O((\varepsilon_{1n})^2)$, where

$$\tau(X) = \frac{1}{K\,(X)} \sum\nolimits_{j=1}^{n} K\,(X, X_j) \times X_j.$$

The point $X - \tau(X)$ lies nearly the linear space $L(X) \approx L_{PCA}(X) = L_g(h(X))$, whence comes the relation

$$(Q_{PCA}(X))^T \times \Delta(X) \approx 0$$

with the error of the order $O((\varepsilon_{1n})^2)$. The latter relation means that the constructed mapping $h(X)$ (84) projects approximately the point X onto the Reconstructed manifold $\boldsymbol{X}_\theta$, whence comes that the embedding $h(X)$ coincides approximately wtth the embedding $h_g(X)$ (15) and the maximum reconstructed error $\delta_\theta(X_0,\ \varepsilon_{1n})$ (14) is $O((\varepsilon_{1n})^2)$.

4) Consider Jacobians $J_{g\bullet h}$ and $J_{h\bullet g}$ of the mappings $r_\theta = g \bullet h$ (12) and $h \bullet g$: $\boldsymbol{Y}_\theta \to \boldsymbol{Y}_\theta$, respectively; the latter mapping $h(g(y))$ is the result of successively applying the reconstruction mapping g to the point $y \in \boldsymbol{Y}_\theta$ and then applying the embedding mapping h to the reconstruction result $g(y) \in \boldsymbol{X}_\theta \subset \boldsymbol{X}_h$. Then the following relations

$$\begin{aligned} J_{g\bullet h}(X) &= \pi(X), \\ J_{h\bullet g}(y)\ &= I_q, \end{aligned}$$

hold true. As a consequence, the residual vector $\Delta(X)$ has null Jacobian, and the following relations

$$r_\theta(X') - r_\theta(X) = X' - X + o(\|X' - X\|),$$
$$h(r_\theta(X')) - h(r_\theta(X)) = h(X') - h(X) + o(\|X' - X\|),$$

hold true for the near points $X, X' \in \boldsymbol{X}$.

5 Numerical Experiments

The GSE-solution $\theta = (h, g)$ gives also a new solution for the ML. To compare the GSE ML-solution with the known methods, the comparative numerical experiments were performed[9]: the GSE-algorithm was compared with the LLE[52], Conformal Eigenmaps[56]; HE[21] (also called Hessian LLE), ISOMAP[63], Landmark ISOMAP[60], LTSA[72].

Two artificial nonlinear Data Manifolds in R^3 were used in the experiments: SwissRoll (Fig. 1(a)) and Spiral (Fig. 2(a)). The training datasets were sampled randomly from these manifolds, and all compared algorithms were applied to the same training samples to construct Embedding and Reconstruction mappings (LLE Reconstruction[53] was used for the algorithms without own reconstruction mapping). The training points on the SwissRoll and Spiral are shown in the Fig. 1(b) and Fig. 2(b), respectively.

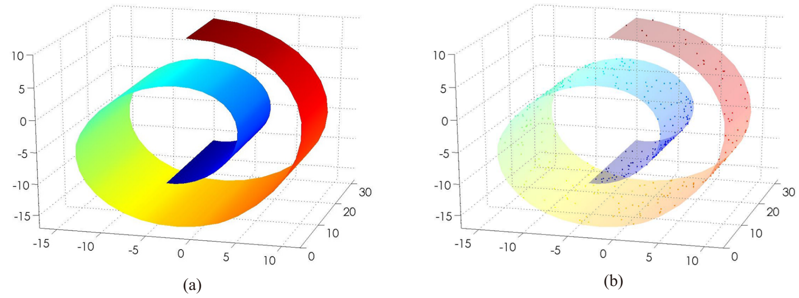


Figure 1. SwissRoll manifold (a) and Training Dataset (b) consisting from $n_{\text{train}} = 450$ points

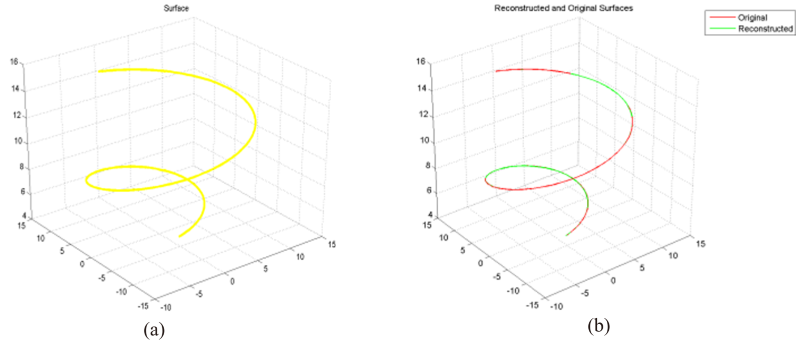
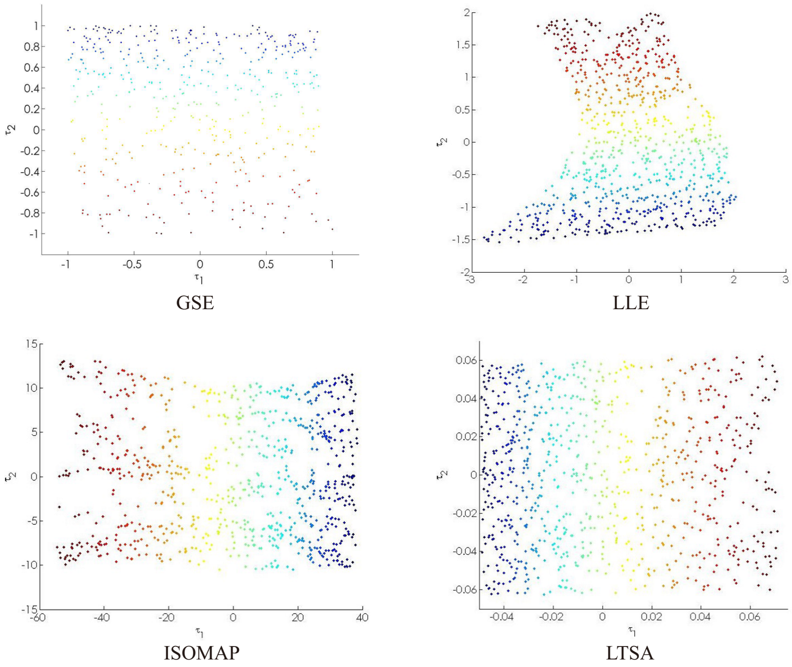


Figure 2. Spiral manifold (a); Original and Solid Reconstructed Spiral manifold (b)

The results of embedding of the training dataset obtaining by the GSE, LLE, ISOMAP and LTSA are shown in the Fig. 3.



GSE

LLE

ISOMAP

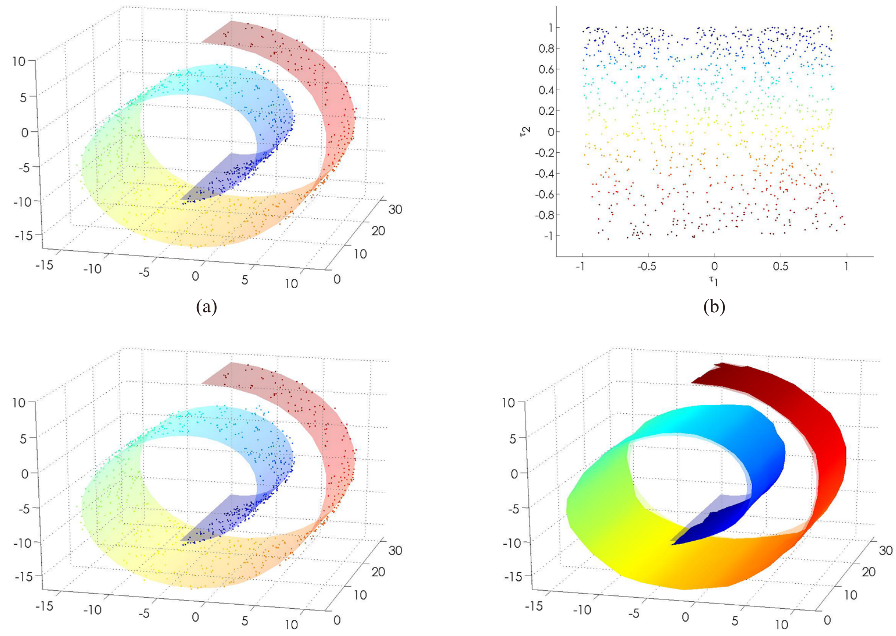
LTSA

Figure 3. Embedding of SwissRoll Training dataset by various methods

The independently generated test dataset from the SwissRoll is shown in the Fig. 4(a). The result of embedding of this dataset obtaining by the GSE is shown in the Fig. 4(b). The result of reconstruction of this dataset is shown in the Fig. 4(c).

The result of "solid reconstruction" of the SwissRoll obtaining by applying the GSE to huge test dataset, which has been generated as the non-random uniform grid on the SwissRoll, is shown in the Fig. 4(d). The result of solid reconstruction of the Spiral obtaining by the GSE is shown in the Fig. 2(b).

The training data sets with various sizes were generated also **o**n the considered artificial manifolds. Then, all the above algorithms were applied to these training dataset to get the Embedding and Reconstruction mappings. After that, the independent large test datasets were sampled randomly from the manifolds and the constructed Embedding and then Reconstruction mappings were applied to the test datasets. The averaged Reconstruction errors (7) based on the test datasets were calculated for each algorithm. These averaged errors are presented in the Fig. 5 (for SwissRole) and Fig. 6 (for Spiral).



(a)

(b)

(c)

(d)

Figure 4. Results for GSE algorithms: (a) Test SwissRoll dataset; (b) Embedding of Test SwissRoll dataset; (c) Reconstruction of Test SwissRoll dataset; (d) Solid reconstruction (based on huge uniform test grid)

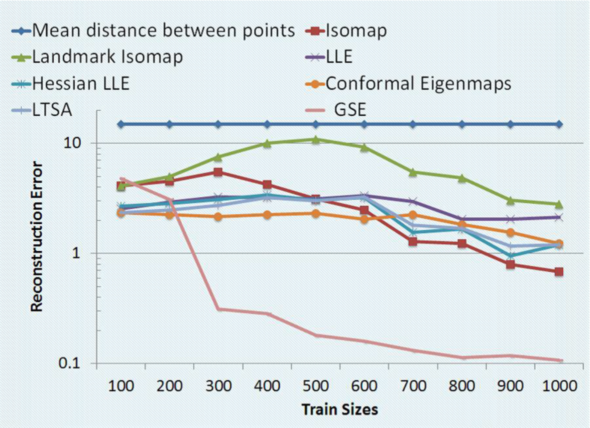


Figure 5. Mean SwissRoll Reconstruction errors calculated for compared methods for various sizes of training dataset

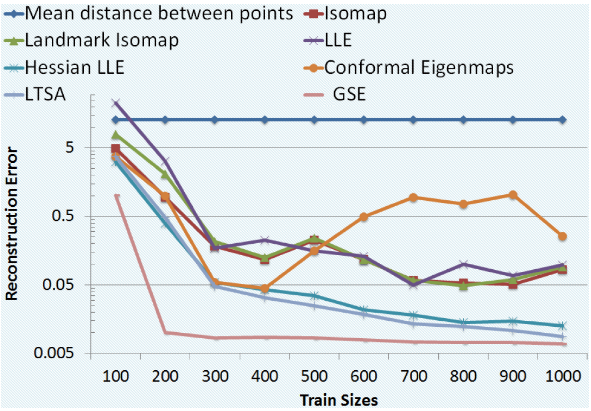


Figure 6. Mean Spiral Reconstruction errors calculated for compared methods for various sizes of training dataset

The Fig. 5 and Fig. 6 show that the proposed GSE algotithm outperforms the compared algorithms with respect to averaged reconstruction error. The numerical experiments performed for other artificial manifolds demonstrate also a good quality of the GSE algorithm.

6 Conclusions

In the paper, we consider Manifold Learning (ML) problems, that is, the Nonlinear Dimensionality Reduction problems under the Manifold Data model, in which a finite dataset is sampled randomly from an unknown low-dimensional Data Manifold embedded in a higher dimensional observation space. A few different formalizations of the ML are discussed. Unlike many works in which the ML is considered as a problem of constructing the low-dimensional parameterization of the Data Manifold, we consider ML as the problem of accurate reconstruction of the Data Manifold from the sample. An accuracy of the Data manifold reconstruction for the Out-of-Sample points characterizes the generalization ability of the ML solution.

We derive asymptotic expansion and local lower and upper bounds for the maximum reconstruction error in a small neighborhood of an arbitrary point from Data Manifold. The expansion and bounds are defined in terms of the distance between tangent spaces to the original Data Manifold and the sample-based Reconstructed Manifold at the selected point and its reconstructed value, respectively. These results imply that the greater the distance between these tangent spaces, the lower the local generalization ability of the ML solution becomes, the poorer the local structure is preserved at the points of the Data Manifold.

By these reasons, we propose an amplification of the ML, called Tangent Bundle Manifold Learning, in which proximity is required not only between the Data Manifold and the sample-based Reconstructed Manifold but also between their tangent spaces. We present a new geometrically motivated Grassman&Stiefel Eigenmaps algorithm that solves this problem, reconstructs accurately the tangent spaces of the Data Manifold and gives a new solution for the ML also. The results of performed comparative numerical experiments are presented.

References

[1] Achlioptas D. Random matrices in data analysis. In: Boulicaut JF, Esposito F, Giannotti F, Pedreschi D, eds. Proc. of the 15th European Conference on Machine Learning. Lecture Notes in Computer Science, 3201. Pisa: Springer Verlag. 2004. 1–8.

[2] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 2003, 15: 1373–1396.

[3] Bengio Y, Delalleau O, Le Roux N, Paiement JF, Vincent P, Ouimet M. Learning eigenfunctions link spectral embedding and kernel PCA. Neural Computation, 2004, 16(10): 2197–2219.

[4] Bengio Y, Delalleau O, Le Roux N, Paiement JF, Vincent P, Ouimet M. Out-of-sample extension for LLE, Isomap, MDS, Eigenmaps, and spectral clustering. In: Thrun S, Saul L, Schölkopf B, eds. Advances in Neural Information Processing Systems, 16. Cambridge, MA: MIT Press. 2004. 177–184.

[5] Bengio Y, Monperrus M. Non-local manifold tangent learning. In: Saul L, Weiss Y, Bottou L, eds. Advances in Neural Information Processing Systems, 17. Cambridge, MA: MIT Press. 2005. 129–136.

[6] Bernstein A, Kuleshov A, Sviridenko Y, Vyshinsky V. Fast aerodynamic model for design technology. Proc. of West-East High Speed Flow Field Conference, WEHSFF-2007. Moscow, Russia: IMM RAS. 2007. http://wehsff.imamod.ru/pages/s7.htm.

[7] Bernstein A, Kuleshov A. Cognitive technologies in the problem of dimension reduction of geometrical object descriptions. Information Technologies and Computer Systems, 2008, 2: 6–19.

[8] Bernstein AV, Burnaev EV, Chernova SS, Zhu F, Qin N. Comparison of three geometric parameterization methods and their effect on aerodynamic optimization. Proc. 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.

[9] Bernstein A, Burnaev E, Erofeev P. Manifold reconstruction in dimension reduction problem. Proc. of the 9^{th} International Conference "Intelligent Information Processing", IIP-2012. Montenegro, Budva: Torus Press. 2012. 196–199.

[10] Brand M. Charting a manifold. In: Becker S, Thrun S, Obermayer K, eds. Advances in Neural Information Processing Systems, 15. Cambridge, MA: MIT Press. 2003. 961–968.

[11] Brand M. From subspace to submanifold methods. In: Hoppe A, Barman S, Ellis T, eds. Proc. of the British Machine Vision Conference. London, UK: BMVA Press, 2004.

[12] Brun A, Westin CF, Herberthson M, Knutsson H. Fast Manifold learning based on Riemannian normal coordinates. In: Kalviainen H, Parkkinen J, Kaarna A, eds. Proc. of the 14th Scandinavian conference on image analysis, SCIA'05, Joensuu, Finland. Lecture Notes in

Computer Science, 3540. Springer, 2005. 920–929.

[13] Bunte K, Biehl M, Hammer B. Dimensionality reduction mappings. IEEE Symposium Series in Computational Intelligence (SSCI) 2011 - Computational Intelligence and Data Mining (CIDM). Paris, France. Piscataway, N.J. IEEE. 2011. 349–356.

[14] Burges CJC. Dimension Reduction: A Guided Tour. Foundations and Trends in Machine Learning, 2010, 2(4): 275–365.

[15] Cayton L. Algorithms for manifold learning. Technical Report[No CS2008–0923]. Univ of California at San Diego (UCSD). Citeseer, June 2005. 541–555.

[16] Chen J, Deng SJ, Huo X. Electricity price curve modeling and forecasting by manifold learning. IEEE Transaction on power systems, 2008, 23(3): 877–888.

[17] Conway JH, Hardin RH, Sloane NJA. Packing lines, planes, etc.: Packing in Grassmannian spaces. Experimental Mathematics, 1999, 5(2): 139–159.

[18] Cox TF, Cox MAA. Multidimensional Scaling. Chapman and Hall, 2001.

[19] Dollár P, Rabaud V, Belongie S. Non-isometric manifold learning: analysis and an algorithm. In: Ghahramani Z, ed. Proc. of the 24th International Conference on Machine Learning. Corvallis, OR, USA. Omni Press. 2007. 241–248.

[20] Dollár P, Rabaud V, Belongie S. Learning to traverse image manifolds. In: Schölkopf B, Platt JC, Hoffman T, eds. Advances in Neural Information Processing Systems, 19. Cambridge. MA: MIT Press. 2007. 361–368.

[21] Donoho DL, Grimes C. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. Proc. of the National Academy of Arts and Sciences, 2003, 100: 5591–5596.

[22] Edelman A, Arias TA, Smith T. The geometry of algorithms with orthogonality constraints. SIAM Journal on Matrix Analysis and Applications, 1999, 20(2): 303–353.

[23] Freedman D. Efficient simplicial reconstructions of manifold from their samples. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2002, 24(10): 1349–1357.

[24] Gisbrecht A, Lueks W, Mokbel B, Hammer B. Out-of-Sample Kernel Extensions for Nonparametric Dimensionality Reduction. In: Proceedings of European Symposium on Artificial Neural Networks, ESANN 2012. Computational Intelligence and Machine Learning. Bruges, Belgium. 2012. 531–536.

[25] Golub GH, Van Loan CF. Matrix Computation. 3rd ed. Baltimore, MD: Johns Hopkins University Press. 1996.

[26] Gorban AN, Kegl B, Wunsch D, Zinovyev AY. Principal Manifolds for Data Visualisation and Dimension Reduction. Springer. Berlin - Heidelberg - New York. 2008.

[27] Gower J, Dijksterhuis GB. Procrustes problems. Oxford University Press. 2004.

[28] Hamm J, Lee DD. Grassmann discriminant analysis: a unifying view on subspace-based learning. Proc. of the 25th International Conference on Machine Learning (ICML 2008). July 2008. 376–383.

[29] He XF, Lin BB. Tangent space learning and generalization. Frontiers of Electrical and Electronic Engineering in China, 2011, 6(1): 27–42.

[30] Hecht-Nielsen R. Replicator neural networks for universal optimal source coding. Science, 1995, 269: 1860–1863.

[31] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507.

[32] Hotelling H. Relations between two sets of variables. Biometrika, 1936, 28: 321–377.

[33] James AT. Normal multivariate analysis and the orthognal group. Ann. Math. Statistics, 1954, 25: 40–75.

[34] Jollie T. Principal Component Analysis. New-York, Springer. 2002.

[35] Izenman AJ. Introduction to manifold learning. Computational Statistics, 2012, 4(5): 439–446.

[36] Karygianni S, Frossard P. Tangent-based manifold approximation with locally linear models. arXiv:1211.1893v1 [cs.LG], 6 Nov. 2012.

[37] Kohonen T. Self-organizing Maps. Springer-Verlag, 3rd Edition, 2000.

[38] Kramer M. Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal, 1991, 37(2): 233–243.

[39] Kriegel H P, Kröger P, Schubert E, Zimek A. A general framework for increasing the robustness

of PCA-based correlation clustering algorithms. Proc. 20th Int. Conf. Scientific and Statistical Database Management (SSDBM). Hong Kong, China, 2008. Lecture Notes in Computer Science, 2008, 5069: 418–435.

[40] Kuleshov A, Bernstein A. Cognitive technologies in adaptive models of complex plants. Information Control Problems in Manufacturing, 2009, 13(1): 1441–1452.

[41] Lafon S, Lee AB. Diffusion maps and coarse-graining: A united framework for dimensionality reduction, graph partitioning and data set parameterization. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2006, 28(9): 1393–1403.

[42] Lee JM. Manifolds and differential geometry. Graduate Studies in Mathematics, 107. Providence: American Mathematical Society, 2009.

[43] Lee JA, Verleysen M. Quality assessment of dimensionality reduction based on k-ary neighborhoods. In: Saeys Y, Liu H, Inza I, Wehenkel L, Van de Peer Y, eds. JMLR Workshop and Conference Proc. Vol. 4: New Challenges for Feature Selection in Data Mining and Knowledge Discovery. Antwerpen, Belgium, 2008. 21–35.

[44] Lee JA, Verleysen M. Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing, 2009, 72(7–9): 1431–1443.

[45] Lee JM. Introduction to Smooth Manifolds. New York, Springer-Verlag. 2003.

[46] Lin T, Zha H, Lee S. Riemannian manifold learning for nonlinear dimensionality reduction. In: Leonardis A, Bischof H, Prinz A, eds. Proc. of ECCV 2006. Berlin Heidelberg, Springer-Verlag: Part I, LNCS 3951. 2006. 44–55.

[47] Ma Y, Fu Y. Manifold Learning Theory and Applications. London: CRC Press, 2011.

[48] Martinetz T, Schulten K. Topology representing networks. Neural Networks, 1994, 7: 507–523.

[49] DeMers D, Cottrell GW. Nonlinear dimensionality reduction. In: Hanson D, Cowan J, Giles L, eds. Advances in Neural Information Processing Systems, 5. San Mateo, CA: Morgan Kaufmann. 1993. 580–587.

[50] Rifai S, Vincent P, Muller X, Glorot X, Bengio Y. Contractive auto-encoders: explicit invariance during feature extraction. In: Getoor L, Scheffer T, eds. Proc. of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, Omnipress. 2011. 833–840.

[51] Rifai S, Dauphin YN, Vincent P, Bengio Y, Muller X. The manifold Tangent Classifier. In: Shawe-Taylor J, Zemel RS, Bartlett P, Pereira F, Weinberger KQ, eds. Advances in Neural Information Processing Systems, 24. Cambridge, MA, MIT Press. 2011.

[52] Saul LK, Roweis ST. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290: 2323–2326.

[53] Saul LK, Roweis ST. Think globally, fit locally: unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research, 2003, 4: 119–155.

[54] Saul LK, Weinberger KQ, Ham JH, Sha F, Lee DD. Spectral methods for dimensionality reduction. In: Chapelle O, Schölkopf B, Zien A, eds. Semisupervised Learning. Cambridge, MA, MIT Press. 2006. 293–308.

[55] Schölkopf B, Smola A, Műller K. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 1998, 10(5): 1299–1319.

[56] Sha F, Saul LK. Analysis and extension of spectral methods for nonlinear dimensionality reduction. In: De Raedt L, Wrobel S, eds. Proc. of the 22nd International Conference on Machine Learning, ICML-05. Bonn, Germany, 2005. ACM International Conference Proceeding Series, 119, New-York NY, ACM. 2005. 785–792.

[57] Silva JG, Marques JS, Lemos JM. A Geometric approach to motion tracking in manifolds. In: Paul MJ, Van Den Hof, Bo W, Weiland S, eds. A Proc. Volume from the 13th IFAC Symposium on System Identification, Rotterdam. 2003.

[58] Silva JG, Marques JS, Lemos JM. Non-linear dimension reduction with tangent bundle approximation. Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP'05). Conference Publications. 4. 85–88.

[59] Silva JG, Marques JS, Lemos JM. Selecting landmark points for sparse manifold learning. In: Weiss Y, Schölkopf B, Platt J, eds. Advances in Neural Information Processing Systems, 18. Cambridge, MA, MIT Press. 2006.

[60] de Silva V, Tehenbaum JB. Global versus local methods in nonlinear dimensionality reduction. In: Becker STS, Obermayer K, eds. Advances in Neural Information Processing Systems, 15. Cambridge, MA: MIT Press, 2003: 705–712.

[61] Song W, Keane AJ. A study of shape parameterisation methods for airfoil optimisation. Proc. of the 10th AIAA / ISSMO Multidisciplinary Analysis and Optimization Conference, AIAA 2004–4482, Albany, New York, American Institute of Aeronautics and Astronautics. 2004.

[62] Strange H, Zwiggelaar R. A generalised solution to the out-of-sample extension problem in manifold learning. Proc. of the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, California, USA: AAAI Press, Menlo Park, California. 2011. 471–478.

[63] Tehenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290: 2319–2323.

[64] Verma N. Mathematical Advances in Manifold Learning. Technical Report. San Diego, University of California. 2008.

[65] Wang J, Zhang ZY, Zha HY. Adaptive manifold learning. In: Saul LK, Weiss Y, Bottou L, eds. Advances in Neural Information Processing Systems, 17. Cambridge, MA, MIT Press. 2005. 1473–1480.

[66] Wang LW, Wang X, Feng JF. Subspace distance analysis with application to adaptive bayesian algorithm for face recognition. Pattern Recognition, 2006, 39(3): 456–464.

[67] Wasserman L. All of Nonparametric Statistics. Berlin, Springer Texts in Statistics. 2007.

[68] Weinberger KQ, Saul LK. Maximum variance unfolding: Unsupervized Learning of image manifolds by semidefinite programming. International Journal of Computer Vision, 2006, 70(1): 77 – 90.

[69] Wolf L, Shashua A. Learning over sets using kernel principal angles. J. Mach. Learn. Res., 2003, 4: 913–931.

[70] Woods RP. Differential geometry of Grassmann manifolds. Proc. Nqt. Acad. Sci. USA, 1967, 57: 589–594.

[71] Yanovich Y, Kuleshov A, Bernstein A. In: Màrkus L, Prokaj V, eds. Asymptotically optimal method in manifold estimation. Abstracts of the XXIX-th European Meeting of Statisticians. Hungary, Budapest. 20-25 July 2013. 25. http://ems2013.eu/conf/upload/BEK086_006.pdf, 2013: 325.

[72] Zhang ZY, Zha HY. Principal manifolds and nonlinear dimension reduction via local tangent Space Alignment. SIAM Journal on Scientific Computing, 2005, 26(1): 313–338.

[73] Zhao DL. Tangential Eigenmaps: A Unifying Geometric for Manifold Learning. Shanghai Jiao Tong University, 2005. http://sites.google.com/site/zhaodeli/paper.