Geometrically Motivated Nonstationary Kernel Density Estimation on Manifold

[1]**Alexander Kuleshov** and [1,2]**Alexander Bernstein** and [1,2]**Yury Yanovich**
[1]Skolkovo Institute of Science and Technology, Moscow, Russia;
[2]Kharkevich Institute for Information Transmission Problems RAS, Moscow, Russia

Abstract

Consider a problem of estimating an unknown high dimensional density whose support lies on unknown low-dimensional data manifold. This problem arises in many data mining tasks, and the paper proposes a new geometrically motivated solution for the problem in manifold learning framework, including an estimation of an unknown support of the density.
Firstly, tangent bundle manifold learning problem is solved resulting in transforming high dimensional data into their low-dimensional features and estimating the Riemannian tensor on the Data manifold. After that, an unknown density of the constructed features is estimated with the use of appropriate kernel approach. Finally, with the use of estimated Riemannian tensor, the final estimator of the initial density is constructed.

The general goal of Data Mining is to extract previously unknown information from a given dataset. Thus, it is supposed that the information is reflected in the structure of a dataset which must be discovered from the data by Data Analysis algorithms. Data mining has a few main "super-problems", which correspond to various tasks: exploratory data analysis, clustering, classification, association pattern mining, outlier analysis, etc. These problems are challenging for Data mining because they are repeatedly used as building blocks in the context of a wide variety of data mining applications (Han and Kamber 2012; Zaki, Mohammed J 2013).

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

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

Although the data for a given data mining problem may have many features, in reality the intrinsic dimensionality of the data support (usually called Data space, DS) of the full feature space may be low. It means that high dimensional data occupy only a very small part in high-dimensional "observation space" whose intrinsic dimension is small. The most popular geometrical data model which describes the low-dimensional structure of the DS is Manifold model (Seung and Lee 2000) by which high-dimensional real-world data lie on or near some unknown low-dimensional Data manifold (DM) embedded in an ambient high-dimensional observation space. Various data analysis problems studied under this assumption about processed data, usually called manifold valued data, are referred to as the Manifold learning problems whose general goal is a discovering of the low-dimensional structure of the high-dimensional DM from given sample (Huo and Smith 2008; Ma and Fu 2011).

Sampling model describes a way for extracting data from the DS. Typically, this model is a probabilistic model: data are selected from the DS independently of each other according to an unknown probability measure on the DS whose support coincides with the DS. Statistical problems for unknown Probabilistic model consist in estimating unknown probability measure or its various characteristics, including density.

Many high dimensional mining algorithms require accurate and efficient density estimators. For example, a general algorithm for scalable mining of patterns in high dimensional spaces approximation is based on low dimensional projections, and this technique is based on a density estimation to determine whether a high dimensional candidate is promising (Müller et al. 2009).

In subspace clustering (Kriegel et al. 2005) and the Pattern Fusion approach for frequent itemset patterns (Zhu et al. 2007) use "step-by-step" so called "jump" algorithms that search the subspace/pattern lattice in a large high dimensional data bases. The density estimation technique (namely, the DensEst one (Müller et al. 2009)), which can be easily integrated into subspace clustering and frequent item set mining algorithms, improves both their efficiency and accuracy.

Classification and clustering are key steps for many data mining tasks whose aim is to discover unknown relationships and/or patterns from large sets of data (Bradley, Fayyad, and Reina 1998). A simple and appealing approach to classification is the K-nearest neighbour method that Finds the K-nearest neighbours of the query point X in the dataset, and then predicts the class label of X as the most frequent one occurring in the K neighbours. However, for large datasets, the time required to compute the neighbourhoods (i.e., the distances of the query from the points in the dataset) becomes prohibitive, making exact answers intractable. Another relevant problem for data mining applications is the approximation of multi-dimensional range queries. Answering range queries, in fact, is one of the simpler data exploration tasks. When the number of dimensions increases, the query time is linear to the size of the dataset (Weber, Schek, and Blott 1998). Thus the problem of efficiently approximating the selectivity of range queries arises naturally. In general, only efficient approximation algorithms can make data exploration tasks in large datasets interactive. Kernel density estimation technique applied to the above tasks our technique allows to efficiently solve range query approximation, classification and clustering problems for very large datasets (Domeniconi and Gunopulos 2004), see also (Zaki, Mohammed J 2013), Ch. 15. To build a clustered index for efficient retrieval of approximate nearest neighbour queries, a density estimation approach was used in (Bennett, Fayyad, and Geiger 1999) to reorganize the data on the disk, with the objective of minimizing the number of cluster scans at query time.

Density-based methods are also used in Outlier analysis (Han and Kamber 2012), Ch. 8, in knowledge discovery algorithms (Stanski and Hellwich 2012), etc. High dimensional density estimation is one of the main techniques for data visualization (Scott 2009; 2012).

There are many methods for estimating an unknown density, including an estimating of high dimensional density. There are some (though a limited number of) methods for estimating an unknown high dimensional density whose support (a domain of density definition) is low-dimensional manifold. The latter methods are known for estimating a density on the known manifold; the obtained estimators are generalized on a case of the unknown manifold in a few papers.

To the best of our knowledge, used techniques give only point density estimators, without estimating a domain of definition of unknown density, and kernel density on unknown manifold estimation problem has not been considered in Manifold learning framework. This framework means that, first, low-dimensional features are constructed with the use appropriate manifold learning technique, and, after this, initial data analysis problem is reduced to a similar problem for constructed features. Finally, the obtained solution to the reduced low-dimensional problem is used for solving the initial high dimensional problem.

The paper presents a new geometrically motivated method for estimating an unknown density on unknown low-dimensional Data manifold based on manifold learning framework. The solution includes an estimation of an unknown support of the unknown density.

Density on manifold estimation: statement and review of related works

Assumptions about Data Manifold

Let $\mathbf{M}$ be an unknown "well-behaved" q-dimensional Data manifold (DM) embedded in an ambient p-dimensional space R^p, $q \leq p$; an intrinsic dimension q is assumed to be known. Assume that the DM $\mathbf{M}$ is a compact manifold with positive condition number (Niyogi, Smale, and Weinberger 2008); thus, no self-intersections, no 'short-circuit.' For simplicity, we assume that the DM is covered by a single coordinate chart φ and, hence, has a form

$$\mathbf{M} = \{X = \varphi(b) \in R^p : b \in \mathbf{B} \subset R^q\} \tag{1}$$

in which chart φ is one-to-one mapping from open bounded Coordinate space $\mathbf{B} \subset R^q$ to the manifold $\mathbf{M} = \varphi(\mathbf{B})$ with inverse map $\psi = \varphi^{-1} : \mathbf{M} \to \mathbf{B}$. Inverse mapping ψ determines low dimensional parameterization on the DM $\mathbf{M}$ (q-dimensional coordinates, or features, $\psi(X)$ of manifold points X), and chart φ recovers points $X = \varphi(b)$ from their features $b = \psi(X)$.

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

$$L(X) = \mathtt{Span}(J_\varphi(\psi(X))) \tag{2}$$

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

As follows from identities $\varphi(\psi(X)) \equiv X$ and $\psi(\varphi(b)) \equiv b$ for all points $X \in \mathbf{M}$ and $b \in \mathbf{B}$, Jacobian matrices $\mathbf{J}_\psi(X)$ and $\mathbf{J}_\varphi(b)$ satisfy the relations

$$\begin{aligned} J_\varphi(\psi(X)) \times J_\psi(X) &\equiv \pi(X); \\ J_\psi(\varphi(b)) \times J_\varphi(b) &\equiv I_q, \end{aligned} \tag{3}$$

in which I_q is $q \times q$ unit matrix and $\pi(X)$ is $p \times p$ projection matrix onto the tangent space $L(X)$ (2) to the DM $\mathbf{M}$ at point $X \in \mathbf{M}$.

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

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

Data Manifold as Riemannian manifold

Let $Z = J_\varphi(\psi(X)) \times z$ and $Z' = J_\varphi(\psi(X)) \times z'$ be vectors from tangent space $L(X)$ with coefficients $z \in R^q$ and $z' \in R^q$ of expansion of these vectors in a basis consisting of columns of Jacobian matrix $J_\varphi(\psi(X))$. An inner product (Z, Z') induced by inner product on R^p equals to $z^T \times \triangle_\varphi(X) \times z$, here $q \times q$ matrix

$$\triangle_\varphi(X) = (J_\varphi(\psi(X)))^T \times J_\varphi(\psi(X)) \tag{4}$$

is metric tensor on the DM $\mathbf{M}$. Thus, $\mathbf{M}$ is Riemannian manifold $(\mathbf{M}, \triangle_\varphi)$ with Riemannian tensor $\triangle_\varphi(X)$ in each manifold point $X \in \mathbf{M}$ smoothly varying from point to point (Jost 2005; Lee 2009). This tensor induces an infinitesimal volume element on each tangent space, and, thus, a Riemannian measure on the manifold

$$m(dX) = \sqrt{|\det \triangle_{\varphi}(X)|} \times d_L(X), \quad (5)$$

where $d_L X$ is a Lebesgue measure on the DM $\mathbf{M}$ induced by exponential mapping $\exp_X$ from the Lebesgue measure on the $L(X)$. Denote $\theta_X(X')$ the volume density function on $\mathbf{M}$ as the square-root of the determinant of the metric $\triangle$ expressed in Riemannian normal coordinates of the point $\exp_X^{-1}(X')$. Strict mathematical definitions of these notations are in (Pennec 1999; Pelletier 2005; Henry, Muoz, and Rodrguez 2013).

Probability measure on Data Manifold

Let $\sigma(\mathbf{M})$ is Borel σ-algebra of $\mathbf{M}$ (the smallest σ-algebra containing all the open subsets of $\mathbf{M}$) and μ is a probability measure on the measurable space $(\mathbf{M}, \sigma(\mathbf{M}))$ whose support coincides with the DM $\mathbf{M}$. Assume that μ is absolutely continuous with respect to the measure m (5), and

$$F(X) = \mu(dX)/m(dX) \quad (6)$$

is its density that separates from zero and infinity uniformly in the $\mathbf{M}$. This measure induces probabilistic measure ν (a distribution of random vector $b = \psi(X)$) on full-dimensional space $\mathbf{B} = \psi(\mathbf{M})$ with standard Borel σ-algebra with density

$$f(b) = d\nu/db = \left|\det_{\varphi}(\varphi(b))\right|^{1/2} \times F(\varphi(b)), \quad (7)$$

with respect to the Lebesgue measure db in R^q. Hence,

$$F(X) = \left|\det_{\varphi}(X)\right|^{-1/2} \times f(\psi(X)). \quad (8)$$

Density on manifold estimation problem

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

Given the dataset $\mathbf{X}_n$, the problem is to estimate the density $F(X)$ (6), including to estimate its support $\mathbf{M}$. An estimation of the DM $\mathbf{M}$ means a construction of q-dimensional manifold $\hat{\mathbf{M}}$ embedded in an ambient Euclidean space R^p which meets manifold proximity property $\hat{\mathbf{M}} \approx \mathbf{M}$ meaning small Hausdorff distance $d_H(\hat{\mathbf{M}}, \mathbf{M})$ between these manifolds. The sought-for estimator $\hat{F}(X)$ defined on the constructed manifold $\hat{\mathbf{M}}$ should provide proximity $\hat{F}(X) \approx F(X)$ for all points $X \in \hat{\mathbf{M}}$.

Therefore, we meet two interrelated topics: estimating a domain of definition (the DM $\mathbf{M}$) of unknown density, which is Manifold learning problems, and estimating an unknown density with manifold support, which is a statistical problem. Next two sections give a short review of related works.

Manifold Learning: Related Works

The goal of Manifold Learning (ML) is to find a description of the low-dimensional structure of an unknown q-dimensional DM $\mathbf{M}$ from random sample $\mathbf{X}_n$ (Freedman 2002). The term "to find a description" is not formalized in general, and it has a different meaning in different articles.

In computational geometry this term means "to approximate (to reconstruct) the manifold": to construct an area $\mathbf{M}^*$ in R^p that is "geometrically" close to the $\mathbf{M}$ in a suitable sense (using some proximity measure between subsets like Hausdorff distance (Freedman 2002)), without finding a low-dimensional parameterization on the DM which usually required in the Machine Learning/Data Mining tasks.

The ML problem in Machine Learning/Data Mining is usually formulated as Manifold embedding problem: given dataset $\mathbf{X}_n$, to construct a low-dimensional parameterization of the DM $\mathbf{M}$ which produces an Embedding mapping

$$h : X \in \mathbf{M} \subset R^p \to y = h(X) \in \mathbf{Y}_h = h(\mathbf{M}) \subset R^q \quad (9)$$

from the DM $\mathbf{M}$ to a Feature Space (FS) $\mathbf{Y}_h$ preserving specific geometrical and topological properties of the DM like local data geometry, proximity relations, geodesic distances, angles, etc. Various Manifold embedding methods such as Linear Embedding, Laplacian Eigenmaps, Hessian Eigenmaps, ISOMAP, etc., are proposed, see the surveys (Huo and Smith 2008; Ma and Fu 2011) and others.

Manifold embedding is usually the first step in various Machine Learning/Data Mining tasks, in which reduced features $y = h(X)$ are used in the reduced learning procedures instead of initial p-dimensional vectors X. If the mapping h preserves only specific properties of high-dimensional data, then substantial data losses are possible when using a reduced vector $y = h(X)$ instead of the initial vector X. To prevent these losses, mapping h must preserve as much available information contained in the high-dimensional data as possible (Freedman 2002); this means the possibility to recover high-dimensional points X from their low-dimensional representations $h(X)$ with small recovery error which can describe a measure of preserving the information contained in high-dimensional data. Thus, it is necessary to find a Recovery mapping

$$g : y \in \mathbf{Y}_h \to X = g(y) \in R^p \quad (10)$$

from the FS $\mathbf{Y}_h$ to the ambient space R^P which, together with the Embedding mapping h (9), ensures proximity

$$r_{h,g}(X) \equiv g(h(X)) \approx X, \ \ \forall X \in \mathbf{M}, \quad (11)$$

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

The reconstruction error $\delta_{h,g}(X) = |X - r_{h,g}(X)|$ is a measure of quality of the pair (h, g) at a point $X \in \mathbf{M}$. This pair determines a q-dimensional Recovered Data manifold (RDM)

$$\mathbf{M}_{h,g} = \{X = g(y) \in R^p : y \in \mathbf{Y}_h \subset R^q\}, \quad (12)$$

embedded in R^p and parameterized by single chart g defined on the FS $\mathbf{Y}_h$. An inequality $d_H(\mathbf{M}_{h,g}, \mathbf{M}) \leq sup_{X \in \mathbf{M}} |r_{h,g}(X) - X|$ implies manifold proximity

$$\mathbf{M} \approx \mathbf{M}_{h,g} \equiv r_{h,g}(\mathbf{M}). \quad (13)$$

There are some (though a limited number of) methods for recovery the DM $\mathbf{M}$ from the FS $\mathbf{Y}_h$. For specific linear manifold, the recovery can be easily found using the Principal Component Analysis (PCA) technique (Jolliffe 2002). For nonlinear manifolds, the sample-based Auto-Encoder Neural Networks (Kramer 1991; Hecht-Nielsen 1995; Berry and Sauer 2017) determine both the embedding and recovery mappings. A general method, which constructs a recovery mapping in the same manner as Locally Linear Embedding algorithm (Singer and Wu 2012) constructs an embedding mapping, has been introduced in (Tyagi, Vural, and Frossard 2013). Manifold recovery based on estimated tangent spaces to the DM $\mathbf{M}$ are used in Local Tangent Space Alignment (Hamm and Lee 2008) and Grassman&Stiefel Eigenmaps (GSE) (Wolf and Shashua 2003) algorithms.

Due to further reasons, Manifold recovery problem can include a requirement to estimate Jacobian matrix J_g of mapping g (10) by certain $p \times q$ matrix $G_g(y)$ providing proximity

$$G_g(y) \approx J_g(y), \ \ \forall y \in \mathbf{Y}_h. \tag{14}$$

This estimator G_g allows estimating the tangent spaces $L(X)$ to the DM $\mathbf{M}$ by q-dimensional linear spaces

$$L_{h,g}(X) = \mathtt{Span}(G_g(h(X))) \tag{15}$$

in R^p which approximates a tangent space to the RDM $\mathbf{M}_{h,g}$ at the point $r_{h,g} \in \mathbf{M}_{h,g}$ and provides tangent proximity

$$L(X) \approx L_{h,g}(X), \ \ \forall X \in \mathbf{M} \tag{16}$$

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

In manifold theory (Jost 2005; Lee 2009), the set composed of manifold points equipped by tangent spaces at these points is called the Tangent bundle of the manifold. Thus, a manifold recovery problem, which includes a recovery of its tangent spaces too, is referred to as the Tangent bundle manifold learning problem: to construct the triple (h, g, G_g) which, additionally to manifold proximity (11), (13), provides tangent proximity (16) (Golub 1996).

Matrix G_g determines $q \times q$ matrix

$$\triangle_{h,g}(X) = G_g^T(h(X)) \times G_g(h(X)) \tag{17}$$

consisting of inner products between columns of the matrix $G\_g(h(X))$ and considered as metric tensor on the RDM $\mathbf{M}_{h,g}$.

Mathematically (Wang, Wang, and Feng 2006), a "preserving the important information of the DM' means that manifold learning algorithm should "recover the geometry" of the DM, and "the information necessary for reconstructing the geometry of the manifold is embodied in its Riemannian metric tensor". Thus, the solution (h, g, G_g) to the tangent bundle manifold learning problem determines Riemannian manifold $(\mathbf{M}_{h,g}, \triangle_{h,g})$ that accurately approximates the Riemannian Data manifold $(\mathbf{M}, \triangle)$.

In real Manifold Learning/Data Mining tasks, intrinsic manifold dimension q is usually unknown too, but this integer parameter can be estimated with high accuracy from given sample (Genovese et al. 2012; Yanovich 2016; 2017; Rozza et al. 2011; Campadelli et al. 2015): an error of dimension's estimator proposed in (Campadelli et al. 2015) has rate $O(\exp(-c \times n))$ in which constant $c > 0$ doesn't depend on sample size n. Because of this, the manifold dimension is usually assumed to be known (or already estimated).

Density estimation: related works

Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables taking values in R^d and having density function $p(x)$. Kernel density estimation is the most widely-used practical method for accurate nonparametric density estimation. Starting with the works of Rosenblatt (Rosenblatt 1956) and Parzen (Parzen 1962), kernel density estimators have the form

$$\hat{p}(x) = \frac{1}{na^d} \sum_{i=1}^{n} K_d \left(\frac{x - X_i}{a} \right), \tag{18}$$

here kernel function $K(t_1, t_2, \ldots, t_d)$ is nonnegative boundedness function that satisfies certain properties the main of which is

$$\int_{R^d} K_d \left(t_1, t_2, \cdots, t_d\right) dt_1 dt_2 \cdots dt_d = 1, \tag{19}$$

and "bandwidth' $a = a_n$ is chosen to approach to zero at a suitable rate as the number n of data points increases. Optimal bandwidth is $a_n = O(n^{-1/(d+4)})$ that yields optimal rate of convergence of Mean Squared Error (MSE) of the estimator $\hat{p}$:

$$\begin{aligned} MSE(\hat{p}) &= \int_{R^d} \left|\hat{p}\left(x\right) - p\left(x\right)\right|^2 p(x)dx \\ &= O(n^{-4/(d+4)}). \end{aligned} \tag{20}$$

Therefore, the use the kernel estimators (18) with MSE of the order $O(n^{-4/(p+4)})$ is not acceptable for high dimensional data.

Various generalization of the estimator (18) was proposed. For example, adaptive kernel estimators were introduced in work (Wagner 1975) in which bandwidth $a = a_n(x)$ in (18) depends on x and is the distance between x and the k-nearest neighbour of x among $X_1, X_2, \ldots, X_n$, and $k = k_n$ is a sequence of non–random integers such that $\lim_{n \to \infty} k_n = \infty$.

There are some works concerning an estimation of an unknown probability density on non-Euclidean spaces such as low-dimensional manifolds, including concrete manifolds (circle, curve, sphere, Grassmann and Stiefel manifolds, etc.).

Density estimators on a general known compact Riemannian manifold without boundary with the use of Fourier expansions technique was proposed in (Hendriks 1990).

The first time, kernel estimators on general known q-dimensional Riemannian manifold embedded in p-dimensional ambient Euclidean space were proposed by Pelletier (Pelletier 2005). Denote $d_{\triangle}(X, X')$ the Riemannian distance (the length of the smallest geodesic curve) between near points X and X' defined by known Riemannian metric tensor $\triangle$. The proposed estimator

$$\hat{p}(x) = \frac{1}{na^q} \sum_{i=1}^{n} \frac{1}{\theta_{X_i}(X)} K_1 \left(\frac{d(X, X_i)}{a} \right), \qquad (21)$$

under the bandwidth $a_n = O(n^{-1/(q+4)})$, has the MSE of the order $O(n^{-4/(q+4)})$ (Pelletier 2005; Henry and Rodriguez 2009) which is acceptable for high dimensional manifold valued data.

The paper (Henry, Muoz, and Rodrguez 2013) generalizes the estimators (21) to the estimators with adaptive kernel bandwidth $a_n(x)$ (similar to the work (Hendriks 1990) for Euclidean space) depending on x.

The estimator (21) assumes that the DM $\mathbf{M}$ is known in advance, and that we have access to certain geometric quantities related to this manifold such as intrinsic distances $d_\triangle(X, X')$ between its points and the volume density function $\theta_X(X')$. Thus, the estimator (21) cannot be used directly in a case where the data lives on an unknown Riemannian manifold of R^p.

The paper (Ozakin and Gray 2009) proposes a more straightforward method that directly estimates the density of the data as measured in the tangent space, without assuming any knowledge of the quantities about the intrinsic geometry of the manifold such as its metric tensor, geodesic distances between its points, its volume form, etc. The proposed estimator

$$\hat{p}(x) = \frac{1}{na^q} \sum_{i=1}^{n} K_1 \left(\frac{d_E(X, X_i)}{a} \right) \qquad (22)$$

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

Asymptotic behavior of kernel density estimators on a compact Riemannian manifold without boundary from a geometry viewpoint was presented in (Park 2012) in which was shown that the asymptotic behavior of the estimators contains a geometric quantity (the sectional curvature) on the unit sphere. This implies that the behavior depends on whether the sectional curvature is positive or negative.

The paper (Kim and Park 2013) generalizes the results of (Park 2012) to a complete Riemannian manifold. Using the fact that kernel function can be defined on the tangent space $L(X)$ $\mathbf{M}$, a new kernel estimator

$$\hat{p}(x) = \frac{1}{na^q \times C(a)} \sum_{i=1}^{n} K_q \left(\frac{1}{a} \exp_X^{-1}(X_i) \right) \qquad (23)$$

is proposed, here constant $C(a) = \frac{1}{a^q} \int_M K_q \left(\frac{1}{a} \exp_X^{-1}(x) \right) m(\mathrm{dX})$ does not depend on $x \in M$ and $C(a) \to 1$ as $a \to 0$. Asymptotic behaviour of the estimator (23) is studied in (Kim and Park 2013) also.

A significant practical limitation of the current density estimation literature is that methods have not been developed for manifolds with boundary, except in simple cases of linear manifolds where the location of the boundary is assumed to be known. This limitation is overcome in (Berry and Sauer 2017) by developing a density estimation method for manifolds with boundary that does not require any prior knowledge of the location of the boundary. A consistent kernel density estimator for manifolds with (unknown) boundary is introduced in (Berry and Sauer 2017) and has the same asymptotic bias in the interior as on the boundary.

The kernel density estimators constructed in Manifold learning framework are proposed in this paper.

Density on manifold estimation: the Solution

Proposed Approach

The proposed approach consists of three stages:

- solving a Tangent bundle manifold learning problem which results in the solution $(h, g, G_g \approx J_g)$;
- estimating a density $f(y)$ of random feature $y = h(X)$ defined on the FS $\mathbf{Y}_h = h(\mathbf{M})$ from feature sample $\mathbf{Y}_n = \{y_i = h(X_i), i = 1, 2, \ldots, n\}$;
- calculating the desired estimator $\hat{F}(X)$ using $f(y)$ and $(h, g, G_g \approx J_g)$.

GSE Solution to the Tangent Bundle Manifold Learning

The solution for Tangent bundle manifold learning is given by the GSE algorithm (Bernstein and Kuleshov 2012; Bernstein, Kuleshov, and Yanovich 2013; Kuleshov and Bernstein 2014; Bernstein, Kuleshov, and Yanovich 2015) and consists of several steps:

1. Apply Local Principal Component Analysis (PCA) to approximate the tangent spaces. $\mathbf{M}$ at points $X \in \mathbf{M}$.
2. Kernel on Manifold definition construction.
3. Tangent Manifold Learning.
4. Embedding mapping construction.
5. Kernel on feature space construction.
6. Constructing the Recovery Mapping and its Jacobian.

Density on the FS Estimation

Under constructed Embedding mapping $h(X)$ from the DM $\mathbf{M}$ to the FS $\mathbf{Y}_h$, unknown probabilistic measure μ on the DM $\mathbf{M}$ with density $F(X)$ (6) gets an unknown probabilistic measure ν on the FS with unknown density defined on full-dimensional space $\mathbf{Y}_h$

$$f(y) = d\nu/dy = \left| \det \left(G_g^T(y) \times G_g(y) \right) \right|^{-1/2} \times F(g(y)); \qquad (24)$$

with respect to Lebesgue measure on R^q.

Consider estimating problem for unknown density $f(y)$ from the feature sample $y_1, y_2, \ldots, y_n$.

Let $k(y, y')$ be chosen kernel on the FS $\mathbf{Y}_h$. Consider kernel estimator for the density $f(y)$ having a form

$$\hat{f}(y) = \frac{1}{n \times C(y)} \sum_{i=1}^{n} k(y, y_i),$$

in which constant $C(y)$ should provide a requirement (19):

$$C(y) = \int_{R^q} k(y, y') f(y') dy',$$

Consider kernel $K(X, X') = k(h(X), h(X'))$ on the **M**, then relation constant can be written

$$C(h(X)) = \int_{\mathbf{M}} K(X, X') \mu(dX'),$$

Assume that kernel $K(X, X')$ is non-zero only for points X' in an asymptotically small a-neighborhood $E_a(X)$, $a \to 0$, of the point X. As was shown in (Yanovich 2016), random sample point $X' \in \mathbf{X}_n$ fallen into the set $E_a(X)$ has conditional asymptotically $(n \to \infty, a \to 0)$ uniform distribution in the intersection of a full dimensional Euclidean a-ball centered at X with the DM **M**; this intersection is close to the q-dimensional a-ball centered at X and lying in the tangent space $L(X)$ (Singer and Wu 2012). Therefore,

$$C(h(X)) \approx \frac{\mu(E_a(X))}{|E_a(X)|} \int_{\mathbf{M}} K(X, X') d_L(X'),$$

where d_L is Lebesgue measure on the DM **M**, $|E_a(X)|$ is the Lebesgue volume of the neighborhood $E_a(X)$, and quantity $\mu(E_a(X))$ can be estimated by a proportion $n(E_a(X))/n$ of sampling points fallen into the neighborhood $E_a(X)$. For example, for the estimator

$$\hat{f}(y) = \frac{1}{n \times a^q \times C(y)} \sum_{i=1}^{n} k_1\left(\frac{d(y, y_i)}{a}\right),$$

constant C equals $\int_{B_q} k_1(|X'|)\, \mathrm{dX}'$ where B_q is unit ball in R^q.

In general case, using exponential mapping $\exp_X$ at point X, write near point X' as $X' = \exp_X(t \times \theta)$, $t \in [0, \infty)$ and $\theta \in S_{q-1}$, and suppose that kernel $K_a(X, X')$, which depends on a small parameter a, has a form $K_a(X, X') = K(X, \theta, t/a)$, $t = |X' - X|$. As it follows from (Yanovich 2017), we have

$$\int_{\mathbf{M}} K_a(X, X') \mu(dX')$$
$$\approx \frac{n_a(X)}{n \times V_q} \times \int_{S_q} \int_0^1 K(X, \theta, t) \times t^{q-1} dt d\theta,$$

here $n_a(X)$ is a number of sampling points fallen into the p-dimensional a-ball centered at X and V_q is the volume of q-dimensional unit ball B_q.

The above formulas allow computing the kernel estimators $\hat{f}(y)$ of an unknown density $f(y)$ on the FS $\mathbf{Y}_h$.

Density on the Manifold Estimation

Based on a representation (8) and estimated Embedding mapping $h(X)$ and Riemannian tensor $\triangle_{h,g}(X)$ (17), the estimator $\hat{F}(X)$ can be computed by the formula

$$\hat{F}(X) = |\det \triangle_{h,g}(X)|^{1/2} \times \hat{f}(h(X)). \quad (25)$$

The approximation $\triangle_{h,g}(X) \approx v^T(X) \times v(X)$, which yields equality $|\det_{h,g}(X)|^{1/2} \approx |det(v(X))|$, allows us to simplify the estimator (25) to formula

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

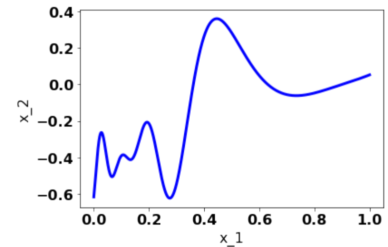

Figure 1: Manifold example.

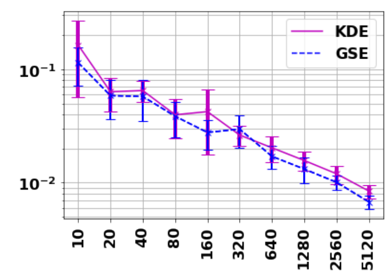

Figure 2: MSE for $\hat{p}$ (KDE, baseline method) and $\hat{F}$ (GSE, proposed method).

Numerical Experiments

The function $x_2 = \sin(30(x_1 - 0.9)^4)\cos(2(x_1 0.9)) + (x_1 - 0.9)/2$, $x_1 \in [0, 1]$, which was used in (Xiong et al. 2007) to demonstrate a drawback of the kernel nonparametric regression (kriging) estimator with stationary kernel (Figure 1), was selected to compare the proposed kernel density estimator $\hat{F}(X)$ (25) and stationary kernel density estimator $\hat{p}(X)$ (22) in R^p. Here $p = 2$, $q = 1$ and $X = (x_1, x_2)^T$. The kernel band-widths were optimized for both methods.

The same training data sets consisting of $n \in \{10, 20, 40, 80, 160, 320, 640, 1280, 2560, 5120\}$ points was used for constructing the estimators; the sample x_1 components were chosen randomly and uniformly distributed on the interval $[0, 1]$. The true probability were calculated theoretically. The errors were calculated for both estimators at the uniform grid on the interval with 100 001 points, then the mean squared errors (MSE) were calculated. Experiments were repeated $M = 10$ times and the mean value of MSE and mean plus/minus standard deviation are shown in Figure 2. The numerical results shows that the proposed approach performs better results than baseline algorithm.

Conclusion

An estimation problem for an unknown density defined on the unknown manifold is solved in Manifold learning framework. A new geometrically motivated solution to this problem is proposed. The algorithm is a geometrically motivated nonstationary kernel density estimator with a single parameter for a kernel width. A numerical experiment with artificial data shows the better results of the proposed approach against ordinary kernel density estimator and could be considered as a proof of concept example.

References

Bennett, K. P.; Fayyad, U.; and Geiger, D. 1999. Density-Based Indexing for Approximate Nearest-Neighbor Queries. In *ACM SIGKDD Conference Proceedings*, 233–243. New York, New York, USA: ACM Press.

Bernstein, A. V., and Kuleshov, A. P. 2012. Tangent Bundle Manifold Learning via Grassmann & Stiefel Eigenmaps.

Bernstein, A.; Kuleshov, A. P.; and Yanovich, Y. 2013. Asymptotically Optimal Method for Manifold Estimation Problem. In *Abstracts of the International Workshop on Statistical Learning*, 8–9.

Bernstein, A.; Kuleshov, A.; and Yanovich, Y. 2015. Information preserving and locally isometric&conformal embedding via Tangent Manifold Learning. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, 1–9. Paris: IEEE.

Berry, T., and Sauer, T. 2017. Density estimation on manifolds with boundary. *Computational Statistics and Data Analysis* 107:1–17.

Bradley, P.; Fayyad, U.; and Reina, C. 1998. Scaling Clustering Algorithms to Large Databases. *KDD* 9–15.

Campadelli, P.; Casiraghi, E.; Ceruti, C.; and Rozza, A. 2015. Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework. *Mathematical Problems in Engineering* 2015:1–21.

Domeniconi, C., and Gunopulos, D. 2004. An Efficient Density-based Approach for Data Mining Tasks. *Knowledge and Information Systems* 10(10):750–770.

Freedman, D. 2002. Efficient simplicial reconstructions of manifolds from their samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(10):1349–1357.

Genovese, C. R.; Perone-Pacifico, M.; Verdinelli, I.; and Wasserman, L. 2012. Minimax Manifold Estimation. *Journal of Machine Learning Research* 13(1):1263–1291.

Golub, G. H. e. C. G. V. L. 1996. *Matrix Computations*.

Hamm, J., and Lee, D. D. 2008. Grassmann Discriminant Analysis : a Unifying View on Subspace-Based Learning. *Proceedings of the 25th international conference on Machine learning* 454(July):376–383.

Han, J., and Kamber, M. 2012. *Data mining : concepts and techniques*. Elsevier.

Hecht-Nielsen, R. 1995. Replicator neural networks for universal optimal source coding.

Hendriks, H. 1990. Nonparametric Estimation of a Probability Density on a Riemannian Manifold Using Fourier Expansions. *The Annals of Statistics* 18(2):832–849.

Henry, G., and Rodriguez, D. 2009. Kernel Density Estimation on Riemannian Manifolds: Asymptotic Results. *Journal of Mathematical Imaging and Vision* 34(3):235–239.

Henry, G.; Muoz, A.; and Rodrguez, D. 2013. Locally adaptive density estimation on Riemannian manifolds. *Statistics and Operations Research Transactions* 37(2):111–130.

Huo, X., and Smith, A. 2008. A Survey of Manifold-Based Learning Methods. *Journal of Machine Learning Research* 1–34.

Jolliffe, I. 2002. *Principal Component Analysis*.

Jost, J. 2005. *Riemannian Geometry and Geometric Analysis*. Springer.

Kim, Y. T., and Park, H. S. 2013. Geometric structures arising from kernel density estimation on riemannian manifolds. *Journal of Multivariate Analysis* 114(1):112–126.

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

Kriegel, H.-P.; Kroger, P.; Renz, M.; and Wurst, S. 2005. A generic framework for efficient subspace clustering of high-dimensional data. *Fifth IEEE International Conference on Data Mining* 8 pp.

Kuleshov, A., and Bernstein, A. 2014. Manifold Learning in Data Mining Tasks. 119–133.

Lee, J. M. 2009. *Manifolds and Differential Geometry (Hardcover)*, volume 107. American Mathematical Society.

Ma, Y., and Fu, Y. 2011. *Manifold Learning Theory and Applications*. London: CRC Press.

Müller, E.; Assent, I.; Krieger, R.; Günnemann, S.; and Seidl, T. 2009. DensEst: Density estimation for data mining in high dimensional spaces. *Proc. SIAM SDM* 175–186.

Niyogi, P.; Smale, S.; and Weinberger, S. 2008. Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete & Computational Geometry* 39(1-3):419–441.

Ozakin, A., and Gray, A. 2009. Submanifold density estimation. *Nips* (2):1–8.

Park, H. S. 2012. Asymptotic behavior of the kernel density estimator from a geometric viewpoint.

Parzen, E. 1962. On Estimation of a Probability Density Function and Mode. In *The Annals of Mathematical Statistics*, volume 33. Institute of Mathematical Statistics. 1065–1076.

Pelletier, B. 2005. Kernel density estimation on Riemannian manifolds. *Statistics & Probability Letters* 73(3):297–304.

Pennec, X. 1999. Probabilities And Statistics On Riemannian Manifolds: Basic Tools For Geometric Measurements. *In IEEE workshop on nonlenear signal and image processing* 194–198.

Rosenblatt, M. 1956. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics* 27(3):832–837.

Rozza, A.; Lombardi, G.; Rosa, M.; Casiraghi, E.; and Campadelli, P. 2011. IDEA: Intrinsic Dimension Estimation Algorithm. 433–442.

Scott, D. W. 2009. *Multivariate density estimation: theory, practice, and visualization*, volume 383.

Scott, D. W. 2012. Multivariate Density Estimation and Visualization. In *Handbook of Computational Statistics*. Berlin, Heidelberg: Springer Berlin Heidelberg. 549–569.

Seung, H. S., and Lee, D. D. 2000. COGNITION: The Manifold Ways of Perception. *Science* 290(5500):2268–2269.

Singer, A., and Wu, H.-T. 2012. Vector diffusion maps and the connection Laplacian. *Communications on Pure and Applied Mathematics* 65(8):1067–1144.

Stanski, A., and Hellwich, O. 2012. A Projection and Density Estimation Method for Knowledge Discovery. *PLoS ONE* 7(10):e44495.

Tyagi, H.; Vural, E.; and Frossard, P. 2013. Tangent space estimation for smooth embeddings of Riemannian manifolds. *Information and Inference* 2(1):69–114.

Wagner, T. J. 1975. Nonparametric Estimates of Probability Densities. *IEEE Transactions on Information Theory* 21(4):438–440.

Wang, L.; Wang, X.; and Feng, J. 2006. Subspace distance analysis with application to adaptive Bayesian algorithm for face recognition. *Pattern Recognition* 39(3):456–464.

Weber, R.; Schek, H. J.; and Blott, S. 1998. A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. *Proceedings of the 24th VLDB Conference* New York C:194–205.

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

Xiong, Y.; Chen, W.; Apley, D.; and Ding, X. 2007. A non-stationary covariance-based Kriging method for metamodelling in engineering design. *International Journal for Numerical Methods in Engineering* 71(6):733–756.

Yanovich, Y. 2016. Asymptotic Properties of Local Sampling on Manifold. *Journal of Mathematics and Statistics* 12(3):157–175.

Yanovich, Y. 2017. Asymptotic Properties of Nonparametric Estimation on Manifold. *JMLR Workshop and Conference Proceedings* 60:1–21.

Zaki, Mohammed J, W. M. J. 2013. *Data Mining and Analysis: Fundamental Concepts and Algorithms*.

Zhu, F.; Yan, X.; Han, J.; Yu, P. S.; and Cheng, H. 2007. Mining Colossal Frequent Patterns by Core Pattern Fusion. In *2007 IEEE 23rd International Conference on Data Engineering*, 706–715. IEEE.