GEOMETRY-ADAPTED GAUSSIAN RANDOM FIELD REGRESSION

Zhen Zhang, Mianzhi Wang, Yijian Xiang, and Arye Nehorai

The Preston M. Green Department of Electrical and Systems Engineering,
Washington University in St. Louis.

ABSTRACT

In this paper, we provide a novel regression algorithm based on a Gaussian random field (GRF) indexed by a Riemannian manifold $(\mathcal{M}, g)$. We utilize both the labeled and unlabeled data sets to exploit the geometric structure of $\mathcal{M}$. We use the recovered heat (H) kernel as the covariance function for the GRF (HGRF). We propose a Monte Carlo integral theorem on Riemannian manifolds and derive the corresponding convergence rate and approximation error. Based on this theorem, we correctly normalize the recovered eigenvector to make it compatible with Riemannian measure. More importantly, we prove that the HGRF is intrinsic to the original data manifold by comparing the pullback geometry and the original geoemtry of $\mathcal{M}$. Essentially it is a semi-supervised learning method, which means the unlabeled data can be utilized to help identify the geometry structure of the unknown manifold $\mathcal{M}$.

Index Terms— regression, manifold learning, semi-supervised learning, Gaussian random field, heat kernel

1. INTRODUCTION

The problem of regression is an important topic in the fields of data analysis and signal processing. The regression model is

$$y_i = f(x_i) + \epsilon_i . i = 1, 2, 3, ..N, \tag{1}$$

where f is an unknown regression function, y_i are observations, x_i belong to $\mathcal{M}$, and $\mathcal{M}$ is an index set. In our paper, we consider the case where $\mathcal{M}$ is a Riemannian manifold embedded in $\mathbb{R}^{\mathrm{n}}$, and ϵ_i is the independent noise during observation. Our goal is to predict f accurately based on the finite labeled data set $\{x_i, y_i\}_{i=1}^N$. Classical methods of regression include linear regression, polynomial regression [1], and wavelet regression [2]. Currently, one of the most popular regression methods is Gaussian process (GP) regression [3] (sometimes called Gaussian random field regression). It assumes that there exists a nonparametric Bayesian prior on f. Predicting $f(x)$ can be regarded as computing the conditional expectation, given observations $\{x_i, y_i\}_{i=1}^N$. So Gaussian process regression is highly related to Kriging in [4]. However, all of the methods mentioned above do not consider the geometry structure of the index set. In fact, as pointed out in [5], the geometry of the data indeed gives a nonparametric prior for statistical inference. Unfortunately, in most practical cases, the geometry of $\mathcal{M}$ is unknown in advance.

Learning the geometry introduces another topic, called manifold learning. In early works, manifold learning aimed at mapping high dimensional data into low dimensional space while preserving the topological or geometric properties of the data. Popular methods include MDS, ISOMAP [6], LLE [7] and SNE [8]. Belkin [9] and Coifman [10] studied the topic from a spectral geometry perspective, making manifold learning a powerful tool for exploring the intrinsic geometry of data. Berry [11] further extended this framework by recovering the geometry coupled with any Riemannian geometry.

Taking advantage of manifold learning, we construct a geometry-adapted Gaussian random field model for regression. The GRF is coupled with the heat kernel of the manifold as its covariance function. Intuitively, the heat kernel is a natural choice since it is positive definite and encodes all the geometry information of the corresponding manifold [12]. In this paper, we prove that the pullback metric induced by GRF is asymptotic to the original metric up to a scalar constant and the heat kernel can guarantee the local smoothness of the regression function from statistical random field view. Further the unlabeled data, which is usually extremely numerous in practice, can also be utilized for identifying the geometry. Generally speaking, more data, whether it is labeled or not, can provide a more reliable geometric prior for regression. From this perspective, our method can be incorporated in a transductive/semi-supervised framework [13]. So our algorithm provides the possibility that we can obtain very accurate prediction even though we do not have much known information. The advantage is in accordance with the big data trend: people produce a tremendous amount of data everyday, however, only a few of them will be labeled. Our simulation results show the significant advantage of the geometry-adapted algorithm, compared with the standard GP method. The results also demonstrate our claim that unlabeled data can help reduce the prediction error. In most cases, the data own its intrinsic geometry structure, such as image [14–16], text data [17], data sampled from stochastic dynamic systems [18–20], audio and video data [21, 22]. So our algorithm will have potential applications in many fields.

2. HEAT KERNEL

In this section, we briefly introduce the heat kernel, and show how to recover it by diffusion maps [10] and correct normalization.

2.1. Learning the Geometry

It's well known that the heat flow along a hypersurface $\mathcal{M}$ (or a Riemannian Manifold) is governed by the heat diffusion equation:

$$\frac{\partial}{\partial t} u = \Delta u \tag{2}$$

$$u(x, 0) = u_0(x), \tag{3}$$

where Δ is the Laplacian-Beltrami operator [23], and u is the temperature distribution on $\mathcal{M}$. The solution of this partial equation is characterized by the heat kernel $p(t, x, y)$: $u(x, t) = \int_{\mathcal{M}} p(t, x, y) u_0(y) \, \mathrm{d}vol(y)$, where the $vol(y)$ is the Riemannian volume measure.

This work was supported by the AFOSR Grant FA9550-16-1-0386 and the ONR Grant N000141310050.

In addition, for compact $(\mathcal{M}, g)$, Sturm-Liouville theorem says that, the heat kernel can be decomposed as [12]

$$p(t,x,y) = \sum_{n=0}^{\infty} \exp(-\lambda_n t)\varphi_n(x)\varphi_n(y), \tag{4}$$

where $\{\varphi_n\}_{n=0}^{\infty}$, which consist of a complete orthonormal basis for $L^2(\mathcal{M}, vol)$, are eigenfunctions of Δ, and $\{\lambda_n\}_{n=0}^{\infty}$ are the corresponding eigenvalues.

Alternatively, the heat kernel is indeed the transition probability of a Brownian motion on $\mathcal{M}$. Since a random walk is the discretized version of the diffusion process, it is possible to learn the geometry of $\mathcal{M}$ by properly formulating the random walk on the data $\{x_i\}_{i=1}^N$ sampled from $\mathcal{M}$. Coifman [10] constructs such a random walk on the data set, and gives a numerical method to approximate Δ:

$$J_\epsilon(x_i, x_j) = \exp(-\|x_i - x_j\|^2/2\epsilon) \tag{5a}$$

$$d_\epsilon(x_i) = \sum_{j=1}^{N} J_\epsilon(x_i, x_j) \tag{5b}$$

$$J_{\epsilon,1}(x_i, x_j) = J_\epsilon(x_i, x_j)/d_\epsilon(x_j) \tag{5c}$$

$$d_{\epsilon,1} = \sum_{j=1}^{N} J_{\epsilon,1}(x_i, x_j) \tag{5d}$$

$$\hat{J}_{\epsilon,1}(x_i, x_j) = J_{\epsilon,1}(x_i, x_j)/d_{\epsilon,1} \tag{5e}$$

$$L_{\epsilon,1} = (I - \hat{J}_{\epsilon,1})/\epsilon. \tag{5f}$$

[10] shows that as $N \to \infty$, $\epsilon \to 0$, we have $L_{\epsilon,1} \to \Delta$. So the eigenvalues $\hat{\lambda}_n$ and eigenvectors $\hat{\varphi}_n$ of $L_{\epsilon,1}$ approximate the eigenvalues and eigenfunctions of Δ respectively.

2.2. Monte Carlo Integral on Manifold

Because of the approximation of eigenvalues and eigenfunctions, we can numerically recover $p(t,x,y)$ of an unknown manifold using the Sturm-Liouville theorem. If we simply normalize $\hat{\varphi}_n$ by the $l2$ norm, the sampling density q will introduce approximation error. That is, the approximated eigenfunctions will be orthonormal in $L^2(\mathcal{M}, \mu)$, where μ is the sampling measure $q(x)vol(x)$, i.e.,

$$\frac{1}{N} = \frac{1}{N}\sum_{i=1}^{N} \hat{\varphi}_n^2(x_i) \to \int_{\mathcal{M}} \hat{\varphi}_n^2(x) q(x)\, \mathrm{d}vol(x). \tag{6}$$

However, as mentioned above, the eigenfunctions should be orthonormal w.r.t. the Riemannian measure:

$$\int_{\mathcal{M}} \hat{\varphi}_n^2(x)\, \mathrm{d}vol(x) = 1. \tag{7}$$

To tackle this problem, we derive the Monte Carlo integral theorem on the Riemannian manifold.

Theorem 1. *Let $\mathcal{M}$ be a compact d-dimensional manifold embedded in R^n without boundary, $f \in C^3(M) \cap L^1(M, vol)$. $\{x_i\}_{i=1}^N$ are i.i.d. random variables sampled from $\mathcal{M}$. We assume that the sampling density $q(x)$ on $\mathcal{M}$ is bounded away from zero and bounded above, i.e., $0 < m = \inf_{x\in\mathcal{M}} q(x) \le \sup_{x\in\mathcal{M}} q(x) = M < \infty$, that $q \in C^3(M) \cap L^1(M, \mathrm{vol})$. We also assume the derivatives of q up to the 2nd order are absolute integrable. Then, for small enough ϵ,*

$$\lim_{N\to\infty} \sum_{i=1}^{N} \frac{(2\pi\epsilon)^{d/2} f(x_i)}{\sum_{j=1}^{N} \exp(-\frac{\|x_i - x_j\|^2}{2\epsilon})} = \int_{\mathcal{M}} f(x)\, \mathrm{d}vol(x) + o(\epsilon) \tag{8}$$

with probability one. (See proof in [24])

Remark 1. The assumption on sampling density $q(x)$ is not quite restrictive, since we can consider q in a Sobolev space.

Corollary 1. *Assume f, q satisfy the conditions mentioned above. Then, with high probability,*

$$\begin{aligned} &\sum_{i=1}^{N} \frac{(2\pi\epsilon)^{d/2} f(x_i)}{\sum_{j=1}^{N} \exp(-\frac{\|x_i - x_j\|^2}{2\epsilon})} \\ &= \int_{\mathcal{M}} f(x)\, \mathrm{d}vol(x) + o(\epsilon, \frac{2\sqrt{G(f,q)}}{\sqrt{N}}, \frac{2\sqrt{M}}{\pi^{-d/4}\sqrt{N}\epsilon^{d/4+1}}), \end{aligned} \tag{9}$$

where $G(f,q)$ is the functional of f, q, defined as

$$G(f,q) = \int_{\mathcal{M}} \frac{f(x)^2}{q(x)}\, \mathrm{d}vol(x) - (\int_{\mathcal{M}} f(x)\, \mathrm{d}vol(x))^2. \tag{10}$$

Remark 2. Here we follow Amit Singer's notation in [25], so $o(\epsilon, \frac{2\sqrt{G(f,q)}}{\sqrt{N}}, \frac{2\sqrt{M}}{\pi^{-d/4}\sqrt{N}\epsilon^{d/4+1}})$ means that there exist positive constants (independent of N and ϵ) C_1, C_2, C_3, such that $|o(\epsilon, \frac{2\sqrt{G(f,q)}}{\sqrt{N}}, \frac{2\sqrt{M}}{\pi^{-d/4}\sqrt{N}\epsilon^{d/4+1}})| \le C_1\epsilon + C_2\frac{2\sqrt{G(f,q)}}{\sqrt{N}} + C_3\frac{2\sqrt{M}}{\pi^{-d/4}\sqrt{N}\epsilon^{d/4+1}}$.

Remark 3. Because of the fact that $\int_{\mathcal{M}} q(x)\, \mathrm{d}vol(x) = 1$, using the Cauchy-Schwarz inequality, we can easily prove $G(f,q) \ge 0$.

Proposition 1. *We normalize the eigenvectors $\hat{\varphi}_n$ by $\tilde{\varphi}_n = \frac{\hat{\varphi}_n}{C_n}$, where*

$$C_n = \sqrt{\sum_{i=1}^{N} \frac{(2\pi\epsilon)^{d/2}\hat{\varphi}_n^2(x_i)}{\sum_{j=1}^{N} \exp(-\frac{\|x_i - x_j\|^2}{2\epsilon})}}, \tag{11}$$

then $\tilde{\varphi}_n$ satisfies $\int_{\mathcal{M}} \tilde{\varphi}_n^2(x)\, \mathrm{d}vol(x) \to 1$.

Proof:

$$\begin{aligned} \int_{\mathcal{M}} \tilde{\varphi}_n^2(x)\, \mathrm{d}vol(x) &\approx \sum_{i=1}^{N} \frac{(2\pi\epsilon)^{d/2}\tilde{\varphi}_n^2(x)}{\sum_{j=1}^{N} \exp(-\frac{\|x_i - x_j\|^2}{2\epsilon})} \\ &= \frac{1}{C_n^2}\sum_{i=1}^{N} \frac{(2\pi\epsilon)^{d/2}\hat{\varphi}_n^2(x)}{\sum_{j=1}^{N} \exp(-\frac{\|x_i - x_j\|^2}{2\epsilon})} \\ &= 1 \end{aligned}$$

We can compute C_n immediately.

Given approximated eigenvalues $\hat{\lambda}_n$ and correctly normalized eigenvectors $\tilde{\varphi}_n$, then we can reconstruct the heat kernel. As we can see, the eigenvalues of the heat kernel $\exp(-\lambda_n t)$ decay at exponential rate, so it is sufficient to select the top M ($M \ll N$) eigenvectors to recover the truncated kernel,

$$p(t,x,y)_{\mathrm{trun}} = \sum_{n=0}^{M} \exp(-\hat{\lambda}_n t)\tilde{\varphi}_n(x)\tilde{\varphi}_n(y). \tag{12}$$

In numerical experiment, we do not need to estimate the dimensionality d of the manifold $\mathcal{M}$, since it is just a constant scalar.

3. PULLBACK GEOMETRY

HGRF can be viewed as an "infinite manifold" embedded in $L^2(\Omega, \mathcal{F}, \mathbb{P})$, with the result that HGRF is naturally coupled with structure. In this section, we derive the pullback metric [26] to show that the heat kernel is intrinsic as a covariance function of GRF. For readers who might not be familiar with random field theory, before formulating our main theorem, we first give some necessary definitions.

Definition 1. *A Gaussian random field f indexed by $\mathcal{M}$ is a collection of Gaussian random variables:*

$$\{f(x) : x \in \mathcal{M}\}, \tag{13}$$

and the finite distributions of $(f_{x_1}, ..., f_{x_n})$ are multivariate Gaussian for each $1 \leq n < \infty$ and each $(x_1, ..x_n) \in \mathcal{M}$.

Definition 2. *Let $(\mathcal{M}, g)$ be a Riemannian manifold, and f be a zero mean Gaussian random field indexed by $\mathcal{M}$. The Riemannian structure induced by f is the pull-back of the standard structure on $L^2(\Omega, \mathcal{F}, \mathbb{P})$, and is given by*

$$g_{\text{pull}}(X_p, Y_p) = E(X_p f Y_p f), \tag{14}$$

where $X_p, Y_p \in T_p\mathcal{M}$.

Theorem 2. *Let $(\mathcal{M}, g)$ be a compact Riemannian manifold, and f be a zero mean Gaussian random field indexed by $\mathcal{M}$. If the covariance function of f is the heat kernel (heat propagation function) on $\mathcal{M}$, then the pullback metric g_{pull} at point $p \in \mathcal{M}$ induced by f is asymptotic to g up to a scalar:*

$$g_{\text{pull}}(p) = \frac{1}{(4\pi t)^{n/2}(2t)}(g_p + t(\frac{1}{6}S_p g_p - \frac{1}{3}Ricci_p) + o(t^2)), \tag{15}$$

where S_p is the scalar curvature and $Ricci_p$ is the Ricci curvature tensor. (See proof in [24])

Remark 4. If $(\mathcal{M}, g)$ is flat, such as a Euclidean space, then the terms in (15) related to curvature will vanish and (15) degenerates to the case in [27]. We also can modify the Riemannian metric on manifold to make it compatible with the task, following the quasi-conformal transformation strategy introduced in [27].

If the prorogation time t is small enough, we immediately see that

$$g_{\text{pull}}(p) \approx \frac{1}{(4\pi t)^{n/2}(2t)} g_p. \tag{16}$$

The constant before g_p does not depend on p, which implies the heat kernel $p(t, x, y)$ approximately preserves the relative distance. Hence, the Riemannian manifold $(\mathcal{M}, g)$ and the Hilbert space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ are "locally isometric" up to a constant (see [24]). The pullback geodesic distance between any two points $p, q \in \mathcal{M}$, is defined as [23]

$$dist_{\text{pull}}(p, q) = \inf_{\gamma(t)} \int_0^1 \sqrt{g_{\text{pull}}(\dot{\gamma(t)}, \dot{\gamma(t)})}\, \mathrm{d}t. \tag{17}$$

(14) and (17) imply that the induced geodesic curve $\tau_{\text{pull}}(t)$ (w.r.t. g_{pull}) is actually the one along which the expected variance of the derivative of f is minimized. Thanks to (16), the geodesic curve $\tau(t)$ (w.r.t. g) approaches $\tau_{\text{pull}}(t)$, and then the smoothness of f along $\tau(t)$ can also be guaranteed.

The metric at any point will change w.r.t. the propagation time t, which reflects the multi-resolution properties of the heat kernel. As pointed out in [28], $p(t, x, y)$ is a family of diffusion wavelets, so our method can be easily generalized into the multi-scale regression framework.

4. HGRF FOR REGRESSION

GRF regression is a state-of-art nonparametric method [3]. GRF is actually a prior over functions: $f \sim \mathcal{GRF}(m, k)$, where m is the mean function (we set it to be zero) and k is the covariance function.

We assume the ϵ_i in regression model (1) are i.i.d. Gaussian noise, so

$$\begin{bmatrix} \boldsymbol{y}^{\text{l}} \\ \boldsymbol{y}^{\text{u}} \end{bmatrix} \Bigg| \, \boldsymbol{x}^{\text{l}}, \boldsymbol{x}^{\text{u}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\mathcal{R}}), \tag{18}$$

$$\boldsymbol{\mathcal{R}} = \begin{bmatrix} \boldsymbol{k}(\boldsymbol{x}^{\text{l}}, \boldsymbol{x}^{\text{l}}) + \sigma^2 \boldsymbol{I} & \boldsymbol{k}(\boldsymbol{x}^{\text{l}}, \boldsymbol{x}^{\text{u}}) \\ \boldsymbol{k}(\boldsymbol{x}^{\text{u}}, \boldsymbol{x}^{\text{l}}) & \boldsymbol{k}(\boldsymbol{x}^{\text{u}}, \boldsymbol{x}^{\text{u}}) \end{bmatrix},$$

where $\boldsymbol{x}^{\text{l}}, \boldsymbol{x}^{\text{u}}$ are labeled and unlabeled data sets respectively, $\boldsymbol{y}^{\text{l}}, \boldsymbol{y}^{\text{u}}$ are function values on labeled and unlabeled data sets respectively, and σ^2 is the variance of noise ϵ_i. The conditional distribution of $\boldsymbol{y}^{\text{u}}$ is,

$$\boldsymbol{y}^{\text{u}} | \boldsymbol{y}^{\text{l}}, \boldsymbol{x}^{\text{l}}, \boldsymbol{x}^{\text{u}} \sim \mathcal{N}(\boldsymbol{\mu}^{\text{u}}, \boldsymbol{\Sigma}^{\text{u}}) \tag{19}$$

where

$$\boldsymbol{\mu}^{\text{u}} = \boldsymbol{k}(\boldsymbol{x}^{\text{u}}, \boldsymbol{x}^{\text{l}})(\boldsymbol{k}(\boldsymbol{x}^{\text{l}}, \boldsymbol{x}^{\text{l}}) + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}^{\text{l}} \tag{20}$$

$$\boldsymbol{\Sigma}^{\text{u}} = \boldsymbol{k}(\boldsymbol{x}^{\text{u}}, \boldsymbol{x}^{\text{u}}) - \boldsymbol{k}(\boldsymbol{x}^{\text{u}}, \boldsymbol{x}^{\text{l}})(\boldsymbol{k}(\boldsymbol{x}^{\text{l}}, \boldsymbol{x}^{\text{l}}) + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}(\boldsymbol{x}^{\text{l}}, \boldsymbol{x}^{\text{u}}). \tag{21}$$

Then, $\boldsymbol{\mu}^{\text{u}}$ is the prediction vector for $\boldsymbol{y}^{\text{u}}$. In our algorithm, k is the heat kernel $p(t, x, y)$ recovered from datasets, which implies our algorithm is also a kind of kernel learning method [29]. The detailed procedure is summarized in Algorithm 1.(Step 2 is the diffusion maps [10] algorithm.)

Algorithm 1 HGRF for regression

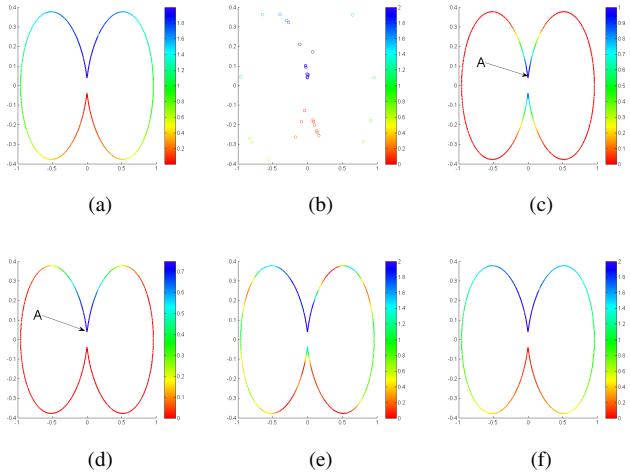
Inputs: Data sets: labeled data $\{\boldsymbol{x}_i^{\text{l}}, \boldsymbol{y}_i^{\text{l}}\}_{i=1}^{n_{\text{l}}}$, and unlabeled data $\{\boldsymbol{x}_i^{\text{u}}\}_{i=1}^{n_{\text{u}}}$. Parameters: number of selected eigenfunctions num_eig, propagation time t, and noise variance σ^2.

Output: The prediction vector $\{\boldsymbol{y}_i^{\text{p}}\}_{i=1}^{n_{\text{u}}}$ for unlabeled data

1. Set $X = [\boldsymbol{x}^{\text{l}}, \boldsymbol{x}^{\text{u}}]$
   $\epsilon = \frac{1}{n_{\text{l}}+n_{\text{u}}} \sum_{i=1}^{n_{\text{l}}+n_{\text{u}}} \min_j \|x_i - x_j\|$
2. (a) Form the affinity matrix $J_{ij} = \{\exp(-\frac{\|x_i - x_j\|^2}{2\epsilon})\}$
   (b) Set $D_i = \sum_{j=1}^{N} J_{i,j}$, $K_{ij} = J_{ij}/D_i D_j$
   (c) Set $Q_i = (\sum_{j=1}^{N} K_{ij})^{\frac{1}{2}}$, $\hat{K}_{ij} = K_{ij}/Q_i Q_j$
   (d) Decompose $[U, S, V] = \text{SVD}(\hat{K})$
   (e) Estimate the eigenvectors $\hat{\varphi}_n = U(:, n)./U(:, 1)$
3. For each n, $\tilde{\varphi}_n = \hat{\varphi}_n / C_n$, define C_n as in (11)
4. Reconstruct the heat kernel as
   $p_{ij} = \sum_{n=1}^{\text{num\_eig}} \exp(-\frac{(1-S_n)t}{\epsilon}) \tilde{\varphi}_n(i) \tilde{\varphi}_n(j)$
5. The prediction vector $\boldsymbol{y}^{\text{p}}$ is
   $\boldsymbol{y}^{\text{p}} = \boldsymbol{p}(\boldsymbol{x}^{\text{u}}, \boldsymbol{x}^{\text{l}})(\boldsymbol{p}(\boldsymbol{x}^{\text{l}}, \boldsymbol{x}^{\text{l}}) + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}^{\text{l}}$

5. NUMERICAL RESULTS

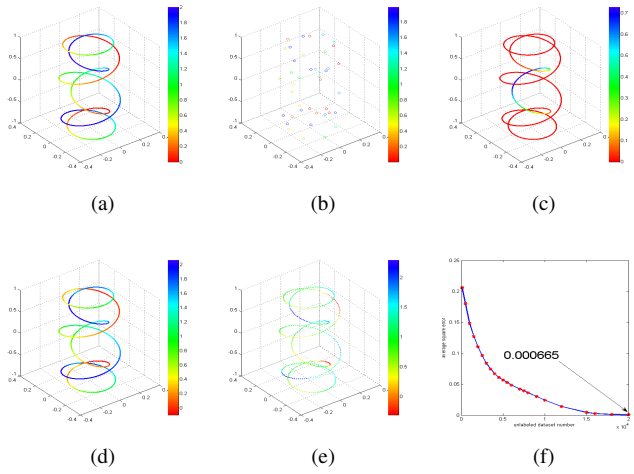
In this section, we will apply our algorithm on both noiseless and noisy data sets. In the first experiment, we compare our algorithm HGRF regression with standard GP regression (here we use GPML toolbox [30]); in the second experiment, we numerically demonstrate our claim that unlabeled data can still help reduce the uncertainty by identifying the data manifold geometry.



(a) (b) (c)

(d) (e) (f)

Fig. 1. (a) The regression function defined on γ; (b) Labeled dataset sampled from γ; (c) Gaussian kernel centered at point A; (d) Heat kernel centered at point A; (e) Regression results with GPML toolbox; (f) Regression results with HGRF



(a) (b) (c)

(d) (e) (f)

Fig. 2. (a) The regression function defined on the helix; (b) Labeled dataset sampled from the helix; (c) Recovered heat kernel with $1.5 * 10^3$ unlabeled data ; (d) Regression results with $1.5 * 10^3$ unlabeled data ; (e) Regression results with 500 unlabeled data; (f) Number of samples–averaged square error curve

5.1. Noiseless Data

Suppose the data are sampled from a closed curve γ:

$$x_1 = (0.5 + 0.46\cos(2\theta))\cos\theta \tag{22}$$

$$x_2 = (0.5 + 0.46\cos(2\theta))\sin\theta. \tag{23}$$

$$\theta \in (0, 2\pi)$$

And the regression function on the curve is: $f = \sin\theta + 1$. We randomly sample 32 points from γ as labeled (or training) dataset, and randomly sample 5000 points from γ as unlabeled dataset. We aim at estimating f on these 5000 points. We set propagation time $t = 0.3$, set regularizer $\sigma^2 = 0.001$ and select the largest 50 eigenvalues and the corresponding eigenvectors to recover the heat kernel.

The regression results are shown in Fig. 1, in which the function value at each point is represented by its color. As we can see in Fig. 1(e) and (f), our algorithm can perfectly recover f. However, the standard GP method with Gaussian kernel as its covariance function fails. We visualize the Gaussian kernel and heat kernel centered at point A in Fig. 1(c) and (d) respectively. For Gaussian kernel, the "information" propagates isotropically into the whole space; however, for the heat kernel, the "information" is constrained along the curves. In fact, this reflects the entropy of GRF [31]. The entropy of the GRF indexed by $\mathbb{R}^2$ is far more than that of the HGRF indexed by γ.

5.2. Noisy Data

We suppose that the datasets are sampled from a helix embedded in $\mathbb{R}^3$, and the we add Gaussian noise on f:

$$x_1 = (0.2 - 0.05\cos(\frac{\pi}{30}t))\sin(\frac{\pi}{5}t) \tag{24}$$

$$x_2 = (0.2 - 0.05\cos(\frac{\pi}{30}t))\cos(\frac{\pi}{5}t) \tag{25}$$

$$x_3 = \sin(\frac{\pi}{30}t) \tag{26}$$

$$t \in (0, 60)$$

The observed function value is $y = \sin(\frac{\pi}{15}t) + 1 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.02^2)$.

We randomly select 40 labeled points. We set propagation time $t = 0.3$, set regularizer $\sigma^2 = 0.001$ and select the largest 500 eigenvalues and the corresponding eigenvectors to recover the heat kernel. We use the average squared error to measure the performance:

$$\text{AVE} = \frac{1}{n_{\text{u}}}\sum_{i=1}^{n_{\text{u}}}(y_i^{\text{p}} - y_i^{\text{u}})^2, \tag{27}$$

where n_{u} is the number of unlabeled data sets, $\boldsymbol{y}^{\text{p}}$ are the prediction values, $\boldsymbol{y}^{\text{u}}$ are the true values on unlabeled data.

This curve is highly twisted as shown in Fig. 2(a), so that it will take many points to accurately "learn" the geometry. As we can see in Fig. 2(d) and Fig. 2(e), our algorithm performs much better on 15000 points than on 500 points. It is clearly shown in Fig. 2(f) that unlabeled data can greatly improve the prediction accuracy**!**

6. CONCLUSION

In this paper, we have developed a geometry-adapted algorithm for regression on unknown Riemannian manifolds. We proposed Monte Carlo integral theorem and derived the explicit relation between the original metric and pullback metric. Our simulation results demonstrated the advantage of the HGRF. The discussion of pullback geometry showed that this algorithm has a great potential for being extended to task-based learning and multi-resolution regression.

7. REFERENCES

[1] H. Theil, "A rank-invariant method of linear and polynomial regression analysis," in *Henri Theils Contributions to Economics and Econometrics*, pp. 345–381, Springer, 1992.

[2] A. Antoniadis, J. Bigot, and T. Sapatinas, "Wavelet estimators in nonparametric regression: a comparative simulation study," *Journal of Statistical Software*, vol. 6, pp. pp–1, 2001.

[3] C. E. Rasmussen, "Gaussian processes for machine learning," 2006.

[4] J.-P. Chiles and P. Delfiner, *Geostatistics: modeling spatial uncertainty*, vol. 497. John Wiley & Sons, 2009.

[5] I. Castillo, G. Kerkyacharian, and D. Picard, "Thomas bayes walk on manifolds," *Probability Theory and Related Fields*, vol. 158, no. 3-4, pp. 665–710, 2014.

[6] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[7] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[8] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in neural information processing systems*, pp. 833–840, 2002.

[9] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[10] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and computational harmonic analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[11] T. Berry and T. Sauer, "Local kernels and the geometric structure of data," *Applied and Computational Harmonic Analysis*, vol. 40, no. 3, pp. 439–469, 2016.

[12] S. Rosenberg, *The Laplacian on a Riemannian manifold: an introduction to analysis on manifolds*. No. 31, Cambridge University Press, 1997.

[13] X. Zhu, "Semi-supervised learning literature survey," 2005.

[14] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[15] C. Bregler and S. M. Omohundro, "Nonlinear image interpolation using manifold learning," *Advances in neural information processing systems*, pp. 973–980, 1995.

[16] G. Mishne and I. Cohen, "Multiscale anomaly detection using diffusion maps," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 111–123, 2013.

[17] K. M. Carter, R. Raich, and A. O. Hero, "Fine: Information embedding for document classification," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1861–1864, IEEE, 2008.

[18] W. Lian, R. Talmon, H. Zaveri, L. Carin, and R. Coifman, "Multivariate time-series analysis and diffusion maps," *Signal Processing*, vol. 116, pp. 13–28, 2015.

[19] R. Talmon and R. R. Coifman, "Empirical intrinsic geometry for nonlinear modeling and time series filtering," *Proceedings of the National Academy of Sciences*, vol. 110, no. 31, pp. 12535–12540, 2013.

[20] R. Talmon, S. Mallat, H. Zaveri, and R. R. Coifman, "Manifold learning for latent variable inference in dynamical systems," *IEEE Transactions on Signal Processing*, vol. 63, no. 15, pp. 3843–3856, 2015.

[21] Y. Keller, R. R. Coifman, S. Lafon, and S. W. Zucker, "Audio-visual group recognition using diffusion maps," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 403–413, 2010.

[22] D. Dov, R. Talmon, and I. Cohen, "Kernel-based sensor fusion with application to audio-visual voice activity detection," *arXiv preprint arXiv:1604.02946*, 2016.

[23] J. M. Lee, *Riemannian manifolds: an introduction to curvature*, vol. 176. Springer Science & Business Media, 2006.

[24] Z. Zhen, W. Mianzhi, X. Yijian, and N. Arye, "Geometry-adapted gaussian random field regression," *In preparation*.

[25] A. Singer, "From graph to manifold laplacian: The convergence rate," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 128–134, 2006.

[26] J. E. Taylor and R. J. Adler, "Euler characteristics for gaussian fields on manifolds," *Annals of Probability*, pp. 533–563, 2003.

[27] S.-i. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, no. 6, pp. 783–789, 1999.

[28] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 21, pp. 7432–7437, 2005.

[29] F. R. Bach, G. R. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *Proceedings of the twenty-first international conference on Machine learning*, p. 6, ACM, 2004.

[30] C. E. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (gpml) toolbox," *Journal of Machine Learning Research*, vol. 11, no. Nov, pp. 3011–3015, 2010.

[31] R. J. Adler and J. E. Taylor, *Random fields and geometry*. Springer Science & Business Media, 2009.