第51卷第5期 中国科学技术大学学报 Vol. 51, No. 5
2021年5月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA May 2021
Received: 2021-01-24; Revised: 2021-04-12 doi:10.52396/JUST-2021-0026

A manifold extended t-process regression

GUO Shiwei, WANG Zhanfeng*, WU Yaohua
Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China
* Corresponding author. E-mail: zfw@ustc.edu.cn

Abstract: A manifold extended t-process regression (meTPR) model is developed to fit functional data with a complicated input space. A manifold method is used to transform covariate data from input space into a feature space, and then an extended t-process regression is used to map feature from feature space into observation space. An estimation procedure is constructed to estimate parameters in the model. Numerical studies are investigated with both synthetic data and real data, and results show that the proposed meTPR model performs well.

Keywords: Gaussian process regression; extended t-process regression; manifold; robustness
CLC number: O212.7 **Document code**: A
2020 Mathematics Subject Classification: 62G08

1 Introduction

A nonparametric regression method, Gaussian process regression (GPR), proposed by Williams and Rasmussen[1] in 1996, is widely used to fit functional data. Rasmussen[2] discussed the detailed algorithm of using the Gaussian process (GP) in the supervised learning of regression and classification, where various covariance functions were proposed and their characteristics were discussed. Shi and Choi[3] introduced methods using the Gaussian process in functional data space. Sun et al.[4] used GPR to predict short-term wind speed, and Liu et al.[5] applied GPR on the prediction of short-term deformation of the tunnel surrounding rock. Many researchers have expanded and improved the Gaussian process from different perspectives. With regard to computational space complexity, Smola and Bartlett[6] used sparse greedy technique to approximate the maximum posterior estimation of the Gaussian process, which performs well when the dataset is large. Seiferth et al.[7] proposed Meta-GP algorithm applying GPR on non-Gaussian likelihood data, which is suitable for data stream processing with low computational complexity. Banerjee et al.[8] proposed a method to solve data storage and processing issues on large dataset by substituting the dataset with a random projection on low dimension subspace. Since GP is susceptible to outliers in data, there are many robust processes proposed to fit functional data. For example, Wauthier and Jordan[9] proposed that GPR tends to overfit in sparse areas of data. They used heavy-tailed stochastic processes to improve the robustness of the estimation. Yu et al.[10] showed that t-process can improve robustness of model. Shah et al.[11] showed that t-process can reduce the overfitting problem while maintaining the excellent properties of GP. Jylänki et al.[12] utilized a t-observation model (Student-t observation model) in GPR and did estimations with expectation propagation to improve robustness of prediction and overcome problems of t-process model. Wang et al.[13] proposed a nonparametric regression method with more robustness than GPR by combining Gaussian process and inverse gamma distribution, which is call extended t-process regression (eTPR).

GPRs and other robust process models are powerful nonparametric regression methods. However, the traditional GPR does not perform well when the dataset is not on vector-space, such as manifold data. For non-smooth data, such as step function, numerical studies show that both GPR and eTPR perform ill. This paper introduces manifold models to devise flexible covariance functions which improved the performance of prediction. Manifolds are now widely used in data processing to change the dimension of the data. When the dimension of the data is large, manifolds can map data to a low-dimensional space to reduce the computational complexity and increase the speed of calculation. Lin and Yao[14] proposed a functional regression method on the manifolds. By means of

Citation: GUO Shiwei, WANG Zhanfeng, WU Yaohua. A manifold extended t-process regression. J. Univ. Sci. Tech. China, 2021, 51(5): 382-389,403.

functional local linear manifold smoothing, the convergence of estimation can reach polynomial speed, and the estimation also performs well on noisy data. Sober et al.[15] used moving least squares to estimate the function on manifolds with linear time complexity, which avoided the non-linear dimensionality reduction process and the loss of information. Zhan and Zhou[16] proposed ManiMIL (manifold based multi-instance learning) and used collapse phenomenon originated from the MIL(multi-instance learning) algorithm to do the prediction, which reduced the calculation time and addressed the collapse issue of LLE (locally linear embedding). In order to enhance the data diversity when reducing data dimension, Gao and Liu[17] reported a method to reconstruct the data with a new defined manifold distance, which improved the recognition rate significantly. Fan and Chen[18] proposed ManiNLR, by combining manifold model with nonlinear regression. They used the manifold model to map high-dimensional space to low-dimensional space, which improved the classification speed. Recently, Calandra et al.[19] combined the manifold method with GPR and created manifold Gaussian process regression (mGPR) by mapping input data to feature space, which improved the accuracy of the prediction, especially at the discontinuous points. Mallasto and Feragen[20] extended GPR to non-vector space by defining wrapped Gaussian processes (WGPs) on Riemannian manifolds.

GPR methods with manifolds, however, are not robust to handle outliers in data. To the best knowledge of authors, there is not a robust process manifold regression model reported in literature. In this paper, we combine t-process and manifold methods to create a robust manifold regression model to fit functional data, which is called the manifold extended t-process regression model (meTPR). We used manifold model to map input space into a feature space. Then the eTPR method is applied to the data in the feature space to capture the nonlinear structures of data. Compared to GPR and eTPR models, the proposed method can fit data from complicated input space, such as non-smooth data. Manifold model significantly improves the accuracy of prediction. In addition, meTPR is more robust than GPR-based manifold methods.

The remainder of the paper is organized as follows. In Section 2, we present the manifold extended t-process regression, and the estimation procedure. Numerical studies and real examples are given in Section 3. Robustness and information consistency properties are showed in Section 4. We conclude in Section 5. Additional technical details and all the proofs are presented in the Appendix.

2 Manifold extended t-process regression

Consider a functional regression model

$$y = F(x) + \epsilon \tag{1}$$

where x is the covariate from input space $\mathscr{X} \subseteq \mathbb{R}^D$, and $y \in \mathscr{Y} \subseteq \mathbb{R}$ is the observation. We focus on the task of learning a regression function $F: \mathscr{X} \to \mathscr{Y}$. To simplify the input space which is usually complicated and improve the accuracy of prediction of non-smooth data, we introduce the manifold model, mapping data space $\mathscr{X}$ to feature space $\mathscr{H}$. Then, we use an eTPR model to depict the relationship between the feature space $\mathscr{H}$ and the output space $\mathscr{Y}$.

The used manifold transformation is a nested mapping as follows,

$$F = f \circ M \tag{2}$$

where $M: \mathscr{X} \to \mathscr{H}$ is the manifold transformation from the input space $\mathscr{X}$ to the feature space $\mathscr{H} \subseteq \mathbb{R}^Q$, and $f: \mathscr{H} \to \mathscr{Y}$ is a function from the feature space $\mathscr{H}$ to the output space $\mathscr{Y} \subseteq \mathbb{R}$. Let $z = M(x) \in \mathscr{H}$ be the features. Then we have $f(z) \in \mathscr{Y}$.

2.1 Manifold transformation

A continuous transformation $M(x) = (T_l \circ \cdots \circ T_1)(x)$ has been used by Calandra et al.[19], where l is the number of layers, x is the input data. Inspired by Calandra et al.[19], we use that transformation in this article. Each T can be written as the following transformation,

$$T_i(x_i) = t(W_i x_i + B_i) \tag{3}$$

where x_i is the input of each layer, $x_1 = x$, t is a transformation function, such as $t(x) = 1/(1+e^{-x})$, and W_i and B_i are the weights and bias of each transformation respectively. For the manifold transformation M, vector θ_M comprises weight parameters and bias parameters of the transformation for each layer, i. e. $\theta_M = [W_1, B_1, \cdots, W_l, B_l]^{\mathrm{T}}$. This transformation can be regarded as one or more widely used coordinate transformations and sigmoid transformations, where the sigmoid transformation is symmetry and has robustness against outliers.

2.2 t-process regression

We now briefly introduce an extended t-process (ETP) and an extended multivariate t-distribution (EMTD). Wang et al.[13] extend a Gaussian process to a t-process using the idea in Reference [21]:

$$f | r \sim \mathrm{GP}(h, rk),\ r \sim \mathrm{IG}(v, \omega) \tag{4}$$

where $\mathrm{GP}(h, rk)$ stands for a GP with a mean function h and a covariance function rk, and $\mathrm{IG}(v, \omega)$ stands for an inverse gamma distribution. Then, f follows an $f \sim \mathrm{ETP}(v, \omega, h, k)$, implying that for any collection of points $x = (x_1, \cdots, x_n)^{\mathrm{T}}$, we have

$$f_n = f(x) = (f(x_1), \cdots, f(x_n))^{\mathrm{T}} \sim \mathrm{EMTD}(v, \omega, h_n, K_n) \tag{5}$$

meaning that f_n has an extended multivariate t-

distribution (EMTD) with the density function,

$$p(z) = |2\pi\omega K_n|^{-1/2}\frac{\Gamma(n/2+v)}{\Gamma(v)}\cdot \left(1+\frac{(z-h_n)^{\mathrm{T}}K_n^{-1}(z-h_n)}{2\omega}\right)^{-(n/2+v)} \quad (6)$$

$h_n=(h(x_1),\cdots,h(x_n))^{\mathrm{T}}$, $K_n=(k_{ij})_{n\times n}$ and $k_{ij}=k(x_i,x_j)$ for some mean function $h(\cdot)\colon \mathscr{X}\to\mathbb{R}$ and covariance kernel $k(\cdot,\cdot)\colon \mathscr{X}\times\mathscr{X}\to\mathbb{R}$.

After mapping the input space to the feature space, we let the eTPR model

$$y(z) = f(z)+\epsilon(z) \quad (7)$$

where z is the feature in feature space $\mathscr{H}$.

We assume f and ϵ are a joint extended t-process (ETP),

$$\begin{pmatrix} f \\ \epsilon \end{pmatrix} \sim \mathrm{ETP}\left(v,\omega,\begin{pmatrix} h \\ 0 \end{pmatrix},\begin{pmatrix} k & 0 \\ 0 & k_\epsilon \end{pmatrix}\right) \quad (8)$$

where h and k are the mean function and kernel function, respectively. The covariance function of ϵ is $k_\epsilon(u,v)=\phi I(u=v)$, where $I(\cdot)$ is an indicative function. We can express the ETP hierarchically as

$$\begin{pmatrix} f \\ \epsilon \end{pmatrix} | r \sim \mathrm{GP}\left(\begin{pmatrix} h \\ 0 \end{pmatrix}, r\begin{pmatrix} k & 0 \\ 0 & k_\epsilon \end{pmatrix}\right) \quad (9)$$

and

$$r\sim \mathrm{IG}(v,\omega) \quad (10)$$

where $\mathrm{IG}(v,\omega)$ is inverse gamma distribution with parameters v and ω. It shows that $y\sim\mathrm{ETP}(v,\omega,h,k+k_\epsilon)$ is joint of $f+\epsilon\,|\,r\sim\mathrm{GP}(h,r(k+k_\epsilon))$ and $r\sim\mathrm{IG}(v,\omega)$, which is the extended t-process regression model (eTPR).

2.3 Estimation

2.3.1 Estimation procedure

Denote the covariate by $x=(x_1,\cdots,x_n)$. The kernel function of eTPR model f is

$$\tilde{k}_{ij} = k(M(x_i),M(x_j)) \quad (11)$$

Let the input data be $\mathscr{D}_n$ and the new data point be u, the model can be written as

$$y(u)\,|\,\mathscr{D}_n \sim \mathrm{EMTD}(n/2+v, n/2+v-1, \tilde{\mu}_n^*, \tilde{\sigma}_n^*+\tilde{s}_0\phi) \quad (12)$$

$$\tilde{\mu}_n^* = E(f(M(u))\,|\,\mathscr{D}_n) = \tilde{k}_u^{\mathrm{T}}\tilde{\Sigma}_n^{-1}y \quad (13)$$

$$\tilde{\sigma}_n^* = \mathrm{Var}(f(M(u))\,|\,\mathscr{D}_n) = s_0(k(M(u),M(u))-\tilde{k}_u^{\mathrm{T}}\tilde{\Sigma}_n^{-1}\tilde{k}_u) \quad (14)$$

where

$$\tilde{\Sigma}_n = \tilde{K}_n+\phi I_n \quad (15)$$

and

$$\tilde{s}_0 = E(r\,|\,\mathscr{D}_n) = \frac{y^{\mathrm{T}}\tilde{\Sigma}_n^{-1}y+2(v-1)}{n+2(v-1)} \quad (16)$$

$\tilde{K}_n$ is the kernel matrix constructed as $\tilde{K}_n=(\tilde{k}_{ij})_{n\times n}$ and $\tilde{k}_u=k(M(x),M(u))$.

2.3.2 Computation

Let $\hat{\theta}=(\hat{\theta}_T,\hat{\theta}_M)$ and $\hat{\phi}$ be the estimated parameters of meTPR process, where $\hat{\phi}$ is the variance parameter of error ϵ, $\hat{\theta}_T$ is parameters of eTPR process f and $\hat{\theta}_M$ is parameters of the manifold model. These parameters can be estimated by minimizing the marginal log likelihood.

Consider the nested mapping $F=f\circ M$. Log marginal likelihood of meTPR is

$$l(\hat{\theta};v) = \sum_{i=1}^{m}\{-\frac{n}{2}\log(2\pi(v-1))-\frac{1}{2}\log|\tilde{\Sigma}_n|-(\frac{n}{2}+v)\log(1+\frac{\tilde{S}}{2(v-1)})+\log(\Gamma(\frac{n}{2}+v))-\log(\Gamma(v))\} \quad (17)$$

where $\tilde{S}=y^{\mathrm{T}}\tilde{\Sigma}_n^{-1}y$. Note that the value of $\tilde{K}_n$ is determined by $\hat{\theta}_T$ and $\hat{\theta}_M$.

The calculation of $\hat{\theta}_T$ is similar to the calculation of parameters in the eTPR[13],

$$\frac{\partial l(\hat{\theta};v)}{\partial\hat{\theta}_{T_k}} = \frac{1}{2}\mathrm{Tr}((\tilde{s}_1\tilde{\alpha}\tilde{\alpha}^{\mathrm{T}}-\tilde{\Sigma}_n^{-1})\frac{\partial\tilde{\Sigma}_n}{\partial\hat{\theta}_{T_k}}) \quad (18)$$

According to the chain rule, we can obtain the gradient-based estimation of θ_M as follows,

$$\frac{\partial l(\hat{\theta};v)}{\partial\hat{\theta}_{M_k}} = \frac{1}{2}\mathrm{Tr}((\tilde{s}_1\tilde{\alpha}\tilde{\alpha}^{\mathrm{T}}-\tilde{\Sigma}_n^{-1})\frac{\partial\tilde{\Sigma}_n}{\partial z}\frac{\partial z}{\partial\hat{\theta}_{M_k}}) \quad (19)$$

where $\tilde{\alpha}=\tilde{\Sigma}_n^{-1}\tilde{y}$, and $\tilde{s}_1=(n+2v)/(2(v-1)+\tilde{S})$. For feature z, $\partial z/\partial\hat{\theta}_{M_k}$ depends only on the input transformation M. A computation procedure for the parameter estimation is as follows,

Step 1. Set initial values of the parameters.

Step 2. For a fixed $\hat{\theta}_M$, update $\hat{\theta}_T$ with (17) and (18).

Step 3. For a fixed $\hat{\theta}_T$, update $\hat{\theta}_M$ with (17) and (19).

Step 4. Repeat Steps 2 and 3 until convergence.

3 Numerical study

This section includes two scenarios of stimulation study, i.e. the step function model and the smooth function model, and compares the performance of the proposed method with those of existing methods. We consider GPR, eTPR, mGPR, and meTPR to fit training data respectively, and obtain the prediction on testing data.

MSE (mean square error),

$$\mathrm{MSE} = \sum_{i=1}^{N}(F_0(x_i^*)-\hat{F}(x_i^*))^2/N$$

and PE (prediction error),

$$\mathrm{PE} = \sum_{i=1}^{N}(y_i^*-\hat{F}(x_i^*))^2/N$$

from each computed method, where $\{(x_i^*,y_i^*)\colon i=1,\cdots,N\}$ are the test data. All simulation results are

based on 100 replications.

3.1 Simulation

3.1.1 Step function

In the first scenario, we consider the following step function model:

$$y = F(x) + w,\ w \sim \mathcal{N}(0, \phi_0^2) \tag{20}$$

where

$$F(x) = \begin{cases} 0, & \text{if } x \leqslant 0; \\ 1, & \text{if } x > 0. \end{cases}$$

Training data points with sample size n are evenly sampled in $[-5,4]$. We take sample size as $n = 20$, 40 and 80, and $\phi_0 = 0.2$ and 0.4. For testing data, 500 data points are generated at equal intervals from $[-5, 5]$. An outlier is set at $(4, 1.5)$. Let

$$M(x) = T(x) = t(Wx + B)$$

be the manifold transformation, where $t(x) = 1/(1+e^{-x})$. Let the dimension of the feature space be 3, W be a 3×1 matrix, and B be a $3\times n$ matrix. Matérn exponential kernel is used,

$$k_m(u, v) = \frac{1}{\Gamma(\alpha)2^{\alpha-1}}(\eta_1 \| u - v \|)^{\alpha}\mathcal{K}_{\alpha}(\eta_1 \| u - v \|) \tag{21}$$

where $\eta_1>0$, $\mathcal{K}_{\alpha}(\cdot)$ is a modified Bessel function of order α, and

$$\mathcal{K}_{\alpha}(x) = \frac{\pi}{2}\frac{I_{-\alpha}(x) - I_{\alpha}(x)}{\sin(\alpha\pi)} \tag{22}$$

and

$$I_{\alpha}(x) = i^{-\alpha}J_{\alpha}(ix) = \sum_{m=0}^{\infty}\frac{1}{m!\,\Gamma(m+\alpha+1)}\left(\frac{x}{2}\right)^{2m+\alpha} \tag{23}$$

Figure 1 shows prediction curves from GPR, eTPR, mGPR, and meTPR based on one simulation dataset. It follows that the meTPR prediction curve fits the indicator function better, compared to GPR and eTPR which ignore the manifold structure. mGPR and GPR are sensitive to outliers, while meTPR shows robustness against the outlier. It is reasonable that meTPR considers both of manifold structure and robustness against ouliers.

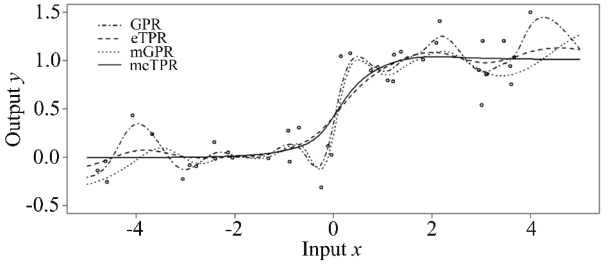

Figure 1. Presents prediction curves from GPR, eTPR, mGPR, and meTPR based on one simulation dataset.

Table 1 shows the mean and standard deviation of the MSE and PE of the predicted results. It shows that meTPR has the smallest MSEs and PEs among the four methods, mGPR is better than GPR and TPR, and eTPR has smaller MSE and PEs than GPR. When sample size becomes larger, MSEs and PEs reduce. It follows that for this non-smooth data, the accuracy of prediction can be improved by the manifold model mapping the input space to the feature space.

3.1.2 Smooth function with outliers

In the second scenario, we consider a smooth function $F(x)$,

$$F(x) = 1/(1 + e^{-3x}) \tag{24}$$

Table 1. MSE, PE and their standard deviation (in parentheses) of GPR, TPR, mGPR and meTPR in Scenario 1.

n	Method	$\phi_0=0.2$		$\phi_0=0.4$	
		MSE mean (SD)	PE mean (SD)	MSE mean (SD)	PE mean (SD)
20	GPR	0.0455(0.0115)	0.0856(0.0122)	0.0648(0.0201)	0.2252(0.0234)
	TPR	0.0421(0.0109)	0.0822(0.0115)	0.0598(0.0183)	0.2205(0.0224)
	mGPR	0.0379(0.0295)	0.0780(0.0300)	0.0590(0.0332)	0.2191(0.0345)
	meTPR	0.0333(0.0110)	0.0734(0.0116)	0.0519(0.0230)	0.2120(0.0249)
40	GPR	0.0279(0.0061)	0.0686(0.0069)	0.0441(0.0114)	0.2049(0.0172)
	TPR	0.0252(0.0055)	0.0658(0.0064)	0.0415(0.0106)	0.2024(0.0168)
	mGPR	0.0206(0.0062)	0.0612(0.0069)	0.0349(0.0098)	0.1956(0.0156)
	meTPR	0.0194(0.0050)	0.0600(0.0061)	0.0346(0.0106)	0.1954(0.0161)
80	GPR	0.0188(0.0044)	0.0590(0.0059)	0.0286(0.0061)	0.1894(0.0126)
	TPR	0.0165(0.0036)	0.0568(0.0053)	0.0279(0.0056)	0.1887(0.0123)
	mGPR	0.0128(0.0074)	0.0530(0.0079)	0.0219(0.0057)	0.1825(0.0122)
	meTPR	0.0118(0.0032)	0.0521(0.0050)	0.0215(0.0054)	0.1822(0.0122)

Table 2. MSE, PE and their standard deviation (in parentheses) of GPR, TPR, mGPR and meTPR in Scenario 2.

n	Method	$\phi_0=0.2$		$\phi_0=0.4$	
		MSE mean (SD)	PE mean (SD)	MSE mean (SD)	PE mean (SD)
20	GPR	0.0259(0.0090)	0.0661(0.0094)	0.0449(0.0212)	0.2064(0.0282)
	TPR	0.0227(0.0081)	0.0630(0.0086)	0.0370(0.0156)	0.1983(0.0222)
	mGPR	0.0205(0.0231)	0.0606(0.0227)	0.0428(0.0486)	0.2050(0.0566)
	meTPR	0.0162(0.0092)	0.0564(0.0094)	0.0356(0.0240)	0.1972(0.0283)
40	GPR	0.0132(0.0048)	0.0537(0.0056)	0.0248(0.0090)	0.1867(0.0150)
	TPR	0.0115(0.0041)	0.0520(0.0051)	0.0227(0.0085)	0.1846(0.0146)
	mGPR	0.0102(0.0196)	0.0510(0.0215)	0.0194(0.0113)	0.1812(0.0154)
	meTPR	0.0086(0.0083)	0.0489(0.0075)	0.0176(0.0089)	0.1795(0.0146)
80	GPR	0.0068(0.0026)	0.0469(0.0038)	0.0144(0.0056)	0.1750(0.0133)
	TPR	0.0057(0.0021)	0.0458(0.0034)	0.0137(0.0051)	0.1743(0.0129)
	mGPR	0.0059(0.0223)	0.0460(0.0219)	0.0104(0.0048)	0.1711(0.0124)
	meTPR	0.0048(0.0083)	0.0450(0.0085)	0.0099(0.0048)	0.1705(0.0123)

Other steps are the same as those for the step function. Tabel 2 shows the mean and standard deviation of the MSE and PE of the predicted results. We obtain the similar conclusion with those for the step function.

3.2 Real data

The proposed meTPR model is applied to dataset for the study of children with Hemiplegic Cerebral Palsy, including 84 girls and 57 boys in primary and secondary schools. These students are divided into two groups ($m=2$): the group playing video games (56%) and the group not playing video games (44%). Average correct rate of Big/Little Circle (BLC) and the average correct rate of Choice Reaction Time (CRT) are measured. More details are in Reference [22]. Before applying the proposed methods, we take logarithm of BLC and CRT mean correct latencies. For GPR, eTPR, mGPR and meTPR, von Mises-inspired kernel was taken.

$$k_{vm}(u,v)=\eta_0\exp(\eta_1(\sum_{l=1}^{p}\cos(u_l-v_l)-p)) \tag{25}$$

where $\eta_0>0$, $\eta_1>0$.

We randomly selected 80% data as the training set and the remaining 20% data as the testing set for calculating the prediction errors under various models. The process is repeated 100 times.

Table 3 shows the mean and standard deviation of prediction errors. It shows that GPR has the largest average prediction error and meTPR has the smallest average prediction error. It follows that meTPR model performs well in improving the accuracy of prediction.

Table 3. Mean and standard deviation of prediction errors using GPR, eTPR, mGPR, and meTPR methods.

Method	Mean(SD) of PE of BLC	Mean(SD) of PE of CRT
GPR	0.1398(0.0299)	0.1630(0.0337)
TPR	0.0361(0.0093)	0.0683(0.0130)
mGPR	0.0611(0.0170)	0.1020(0.0244)
meTPR	0.0333(0.0628)	0.0475(0.0102)

4 Robustness and information consistency

4.1 Robustness

The manifold extended t-process regression can provide estimation with greater robustness than mGPR when data includes outliers. Let $\widehat{\theta}=(\widehat{\theta}_T,\widehat{\theta}_M)$ and

$\widehat{F}_T(u)=\widehat{\mu}_n^*=\mu_n^*\mid_{\theta=\widehat{\theta}}$, $V_T=\widehat{\sigma}_n^*=\sigma_n^*\mid_{\theta=\widehat{\theta}}$

be the predicted mean and variance of $F(u)$ in meTPR. $\widehat{F}_G(u)$ and V_G are the predictions in the manifold Gaussian process. Let $M_T=(\widehat{F}_T(u)-F_0(u))/\sqrt{V_T}$ and $M_G=(\widehat{F}_G(u)-F_0(u))/\sqrt{V_G}$ be two t-test statistics for a null hypothesis $F(u)=F_0(u)$. When the kernel function is bounded, if $y_j\to\infty$ for some j, then $M_G\to\infty$, while M_T is still bounded. Then M_T for meTPR is more robust against outliers compared to that for mGPR.

Let $T(F_n)=T_n(y_1,\cdots,y_n)$ be an estimation of θ, where F_n is the empirical distribution of $\{y_1,\cdots,y_n\}$, and T is functional on the distributions. The influence function of T on F is defined as

$$\mathrm{IF}(y;T,F)=\lim_{t\to 0}\frac{T((1-t)F+t\delta_y)-T(F)}{t} \tag{26}$$

where δ_y is 1 on point y, 0 on other points. The influence function can show the degree of change of the estimated parameter after adding a disturbance to the data set, then it can reflect the robustness of the estimation method. For the meTPR model, we have the following proposition.

Proposition 4.1 Assume the kernel function k is bounded continuous differentiable on θ, then for a given v, the estimated parameters of meTPR, $\widehat{\theta}$, has a bounded influence function, while that from the mGPR does not.

4.2 Information consistency

Let $p_{\phi_0}(y|F_0, x)$ be the density function to generate the data y given x in true model $y=F_0(x)+\epsilon$, where F_0 is the true F. Let $p_\theta(F)$ be a measurement of the random process F on space $\mathscr{F}=\{F(\cdot): \mathscr{X}\to\mathbb{R}\}$. Let

$$p_{\phi,\theta}(y \mid x) = \int_{\mathscr{F}} p_\phi(y \mid F, x)\,\mathrm{d}p_\theta(F) \tag{27}$$

be the density function to generate the data y given x under the meTPR model. In this case, meTPR model is different from true model. Let ϕ_0 be the true value of ϕ. Let p_{ϕ_0}, $\widehat{\theta}(y|x)$ be the estimated density function of meTPR model. Denote

$$D[p_1, p_2] = \int(\log p_1 - \log p_2)\,\mathrm{d}p_1$$

by the Kullback-Leibler distance between two densities p_1 and p_2. According to Seeger et al.[23], if $E_x(D[p_{\phi_0}(y \mid F_0, x), p_{\phi_0,\hat{\theta}}(y \mid x)]) \to 0$ as $n \to \infty$, then we call meTPR model information consistent, which is presented in the following proposition.

Before presenting the information consistency of the meTPR, we briefly introduce a reproducing kernel Hilbert space[24]. Assume $\mathscr{F}$ is a Hilbert space of functions $F: \mathscr{X}\to\mathbb{R}$ with an inner product $\langle\cdot,\cdot\rangle_{\mathscr{F}}$. We call $\mathscr{F}$ a reproducing kernel Hilbert space associated with a kernel function $\tilde{k}$, where the kernel function $\tilde{k}: \mathscr{X}\times\mathscr{X}\to\mathbb{R}$ satisfies

① $\forall x \in \mathscr{X}$, $\tilde{k}(\cdot, x) \in \mathscr{F}$;

② $\forall x \in \mathscr{X}$, $\forall F \in \mathscr{F}$,

$$\langle F, \tilde{k}(\cdot, x)\rangle_{\mathscr{F}} = \delta_x(F) = F(x).$$

Proposition 4.2 Under the appropriate conditions in Lemma A.1 and condition that $\| F_0 \|_k$ is bounded and $E_x(\log|I_n+\phi_0^{-1}\tilde{K}_n|)=o(n)$ holds, we have

$$\frac{1}{n}E_x(D[p_{\phi_0}(y \mid F_0, x), p_{\phi_0,\hat{\theta}}(y \mid x)]) \to 0, \text{ as } n \to \infty \tag{28}$$

where the expectation is taken over the distribution of x and $\| F_0 \|_k$ is norm of F_0 in the reproducing kernel Hilbert space associated with the kernel function $\tilde{k}(\cdot, \cdot; \theta)=k(M(\cdot), M(\cdot); \theta)$.

5 Conclusions

In order to solve the difficulty of fitting with outliers and in complicated covariate space, we proposed a manifold t-process regression (meTPR) model. We proposed a parameter estimation method and studied the theoretical properties of the model. The proposed model is robust to outliers, and performs well for non-smooth and complicated covariate space. Although Y is one-dimensional in this article, the model can be extended to multi-dimensional dependent functional data.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (11971457), the Anhui Provincial Natural Science Foundation (1908085MA06) and the Fundamental Research Funds for the Central Universities (WK2040000035).

Conflict of interest

The authors declare no conflict of interest.

Author information

GUO Shiwei is currently a master student under the tutelage of Assoc. Prof. Wang Zhanfeng at University of Science and Technology of China. His research interests focus on functional data.

WANG Zhanfeng (corresponding author) received his PhD degree from University of Science and Technology of China (USTC). He is currently an associate professor at USTC. His research interests focus on functional data analysis and biostatistics.

References

[1] Williams C, Rasmussen C. Gaussian processes for regression. Advances in Neural Information Processing Systems, 1995, 8: 514-520.

[2] Rasmussen C E. Gaussian processes in machine learning. In: Summer School on Machine Learning. Berlin: Springer, 2004: 63-71.

[3] Shi J Q, Choi T. Gaussian Process Regression Analysis for Functional data. Boca Raton, FL: CRC Press, 2011.

[4] Sun B, Yao H, Liu T. Short-term wind speed forecasting based on Gaussian process regression model. Proceedings of the Chinese Society for Electrical Engineering, 2012, 32(29): 104-109.

[5] Liu K Y, Fang Y, Liu B G, et al. Intelligent deformation prediction model of tunnel surrounding rock based on genetic-Gaussian process regression coupling algorithm. Journal of the China Railway Society, 2011, 33: 101-106. (In Chinese)

[6] Smola A J, Bartlett P L. Sparse greedy Gaussian process regression. In: Advances in Neural Information Processing Systems 13. Cambridge, MA: MIT Press, 2001: 619-625.

[7] Seiferth D, Chowdhary G, Mühlegg M, et al. Online Gaussian process regression with non-Gaussian likelihood. In 2017 American Control Conference (ACC). IEEE,

2017: 3134-3140.
[8] Banerjee A, Dunson D B, Tokdar S T. Efficient Gaussian process regression for large datasets. Biometrika, 2013, 100: 75-89.
[9] Wauthier F L, Jordan M I. Heavy-tailed process priors for selective shrinkage. In: Advances in Neural Information Processing Systems 23. Cambridge, MA: MIT Press, 2010: 2406-2414.
[10] Yu S, Tresp V, Yu K, et al. Robust multi-task learning with *t*-processes. In: Proceedings of the 24th International Conference on Machine learning. New York: Association for Computing Machinery, 2007: 1103-1110.
[11] Shah A, Wilson A, Ghahramani Z. Student-*t* processes as alternatives to Gaussian processes. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics. Cambridge, MA: PMLR, 2014: 877-885.
[12] Jylänki P, Vanhatalo J, Vehtari A. Robust Gaussian process regression with a student-*t* likelihood. Journal of Machine Learning Research, 2011, 12: 3227-3257.
[13] Wang Z, Shi J Q, Lee Y. Extended *t*-process regression models. Journal of Statistical Planning and Inference, 2017, 189: 38-60.
[14] Lin Z, Yao F. Functional regression on manifold with contamination. https://arxiv.org/abs/1704.03005.
[15] Sober B, Aizenbud Y, Levin D. Approximation of functions over manifolds: A moving least-squares approach. https://arxiv.org/abs/1711.00765.
[16] Zhou Zhihua, Zhan Dechuan. A manifold learning-based multi-instance regression algorithm. Chinese Journal of Computers, 2006, 29(11): 1948-1955. (In Chinese)
[17] Gao Y, Liu Y J. Diversity based discriminant muti-manifold learning for dimensionality reduction. Automation and Instrumentation, 2020(4): 30-34. (In Chinese)
[18] Fan J F, Chen D C. Combining manifold learning and nonlinear regression for head pose estimation. Journal of Image and Graphics, 2012, 17(8): 1002-1010. (In Chinese)
[19] Calandra R, Peters J, Rasmussen C E, et al. Manifold Gaussian processes for regression. In: 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016: 3338-3345.
[20] Mallasto A, Feragen A. Wrapped Gaussian process regression on Riemannian manifolds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 5580-5588.
[21] Lee Y, Nelder J A. Double hierarchical generalized linear models (with discussion). Journal of the Royal Statistical Society: Series C (Applied Statistics), 2006, 55: 139-185.
[22] Xu P, Lee Y, Shi J Q, et al. Automatic detection of significant areas for functional data with directional error control. Statistics in Medicine, 2019, 38: 376-397.
[23] Seeger M W, Kakade S M, Foster D P. Information consistency of nonparametric Gaussian process methods. IEEE Transactions on Information Theory, 2008, 54: 2376-2382.
[24] Berlinet A, Thomas-Agnan C. Reproducing Kernel Hilbert Spaces in Probability and Statistics. Berlin: Springer Science & Business Media, 2011.
[25] Hampel F R, Ronchetti E M, Rousseeuw P J, et al. Robust Statistics: The Approach Based on Influence Functions. Hobooken, NJ: Wiley, 2011.

流形拓展 t-过程回归

郭世威,王占锋*,吴耀华
中国科学技术大学管理学院统计与金融系,安徽合肥 230026
*通讯作者. E-mail:zfw@ustc.edu.cn

摘要: 本文提出了一种流形拓展 t-过程回归模型,用来分析带有复杂协变量的函数型数据. 该流形拓展 t-过程回归模型可将协变量数据变换至特征空间,然后用拓展 t-过程回归将特征空间数据转换到观测数据空间,从而对观测数据进行建模. 我们建立了一个估计程序来估计模型中的参数. 对真实数据和模拟数据进行了分析,结果说明所提流形拓展 t-过程回归模型是可行的.

关键词: 高斯过程回归;拓展 t-过程回归;流形;稳健性

Appendix

Lemma A. 1 Under meTPR model (1), assume that the covariance kernel function k is bounded and continuous on the parameter θ, and $\widehat{\theta}$ converges to θ when $n\to\infty$. Then, for a positive constant c and any $\varepsilon>0$, when n is large enough, we have

$$\frac{1}{n}(-\log p_{\phi_0,\widehat{\theta}}(y\mid x)+\log p_{\phi_0}(y\mid F_0,x))\leqslant\frac{1}{n}\{\frac{1}{2}\log|I_n+\phi_0^{-1}\widetilde{K}_n|+\frac{q^2+2(v-1)}{2(n+2v-2)}(\|F_0\|_k^2+c)+c\}+\varepsilon \tag{A1}$$

where $\widetilde{K}_n=(k(M(x_i),M(x_j)))_{n\times n}$, $q^2=(y-F_0(x))^{\mathrm{T}}(y-F_0(x))/\phi_0$, and I_n is the $n\times n$ identity matrix. $\|F_0\|_k$ is the reproducing kernel Hilbert space norm of F_0 associated with kernel function $k(M(\cdot),M(\cdot);\theta)$.

Proof Assume r is a random variable following inverse gamma distribution IG$(v,(v-1))$. let GP(h,k) be Gaussian process with mean function h and covariance function k. Conditional on r, we have

$$\begin{pmatrix}F\\ \epsilon\end{pmatrix}\mid r\sim\mathrm{GP}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}rk&0\\0&rk_\epsilon\end{pmatrix}\right) \tag{A2}$$

Then conditional on r, the extended t-process regression model

$$y(x)=F(x)+\epsilon \tag{A3}$$

becomes Gaussian process regression model

$$y(x)=\widetilde{F}(x)+\widetilde{\epsilon} \tag{A4}$$

where $\widetilde{F}=F\mid r\sim\mathrm{GP}(0,rk(M(\cdot),M(\cdot);\theta))$, $\widetilde{\epsilon}\mid r\sim\mathrm{GP}(0,rk_\epsilon(M(\cdot),M(\cdot);\phi_0))$, and $\widetilde{F}$ and error term $\widetilde{\epsilon}$ are independent. Denoted the computation of conditional probability density for given r by $\widetilde{p}$. For the model $y(x)=\widetilde{F}(x)+\widetilde{\epsilon}$, let

$$p_G(y\mid r,x)=\int_{\mathscr{F}}p_{\phi_0}(y\mid\widetilde{F},r,x)\mathrm{d}\widetilde{p}_\theta(\widetilde{F}) \tag{A5}$$

$$p_0(y\mid r,x)=p_{\phi_0}(y\mid F_0,r,x) \tag{A6}$$

where $\widetilde{p}_\theta$ is the induced measure from Gaussian process GP$(0,rk(M(\cdot),M(\cdot);\widehat{\theta}))$. We know that variable r is independent of covariates x. Easily, we show that

$$p_{\phi_0,\widehat{\theta}}(y\mid x)=\int p_G(y\mid r,x)\,g(r)\mathrm{d}r \tag{A7}$$

$$p_{\phi_0}(y\mid f_0,x)=\int p_0(y\mid r,x)g(r)\mathrm{d}r \tag{A8}$$

Suppose that we have

$$-\log p_G(y\mid r,x)+\log p_0(y\mid r,x)\leqslant\frac{1}{2}\log|I_n+\phi_0^{-1}\widetilde{K}_n|+\frac{r}{2}(\|F_0\|_k^2+c)+c+n\varepsilon \tag{A9}$$

for any given r. So we get

$$-\log\int p_G(y\mid r,x)g(r)\mathrm{d}r\leqslant\frac{1}{2}\log|I_n+\phi_0^{-1}\widetilde{K}_n|+c+n\varepsilon-\log\int p_0(y\mid r,x)\exp\{-(\frac{r}{2}(\|F_0\|_k^2+c))\}g(r)\mathrm{d}r \tag{A10}$$

Let $g^*(r)$ be the density function of IG$(v+n/2,(v-1)+q^2/2)$. It easily shows that

$$\int p_0(y\mid r,x)\exp\{-(\frac{r}{2}(\|F_0\|_k^2+c))\}g(r)\mathrm{d}r=\int p_0(y\mid r,x)g(r)\mathrm{d}r\int\exp\{-(\frac{r}{2}(\|F_0\|_k^2+c))\}g^*(r)\mathrm{d}r \tag{A11}$$

We have

$$\begin{aligned}&-\log p_{\phi_0,\widehat{\theta}}(y\mid x)+\log p_{\phi_0}(y\mid F_0,x)\leqslant\\&\frac{1}{2}\log|I_n+\phi_0^{-1}\widetilde{K}_n|+c-\log\int\exp\{-(\frac{r}{2}(\|F_0\|_k^2+c))\}g^*(r)\mathrm{d}r\leqslant\\&\frac{1}{2}\log|I_n+\phi_0^{-1}\widetilde{K}_n|+c+\frac{\|F_0\|_k^2+c}{2}\int rg^*(r)\mathrm{d}r=\\&\frac{1}{2}\log|I_n+\phi_0^{-1}\widetilde{K}_n|+\frac{q^2+2(v-1)}{2(n+2v-2)}(\|F_0\|_k^2+c)+c+n\varepsilon\end{aligned} \tag{A12}$$

which shows that Lemma A. 1 holds.

(Continued on p. 403)

$$\widetilde{S}_1(\boldsymbol{x}) = f(\boldsymbol{x})\boldsymbol{I}_{p+1} + O_p(n^{-\alpha}),$$

so we have

$$\{\widetilde{S}_1(\boldsymbol{x})\}^{-1} = f^{-1}(\boldsymbol{x})\boldsymbol{I}_{p+1} + O_p(n^{-\alpha}).$$

Similar to the proof of the Theorem 2.1, by Chebyshev's inequality and the Lyapunov CLT, we have

$$n^{(1-p\alpha)/2}(\widetilde{M}(\boldsymbol{x}) - M(\boldsymbol{x}) - B_{ll}(\boldsymbol{x})) \xrightarrow{\mathrm{d}} N(o, \Sigma),$$

where

$$B_{ll}(\boldsymbol{x}) = \frac{c^2 n^{-2\alpha} f(\boldsymbol{x})}{2(1-2\alpha)} \begin{pmatrix} \mathrm{tr}\{\mathscr{H}_m(\boldsymbol{x})\} \\ \boldsymbol{0}_{p\times 1} \end{pmatrix}, \quad \Sigma = \frac{\sigma_\varepsilon^2}{c^p(1+p\alpha)f(\boldsymbol{x})} \begin{pmatrix} R_2(K) & \boldsymbol{0}_{1\times p} \\ \boldsymbol{0}_{p\times 1} & \widetilde{R}_2(K) \end{pmatrix},$$

Since one can easily verify that the conditions in Lemma A.1 are all satisfied. Therefore, we can apply the Liapounov CLT to conclude Theorem 3.2.

Proof of Corollary 3.2 The proof of Corollary 3.2 is analogous to that of Corollary 3.1. To avoid duplication, descriptions are not provided in this paper.

(Continued from p. 389)

Proof of Proposition 4.1 The score function of θ_k in the meTPR model is

$$s_k(\theta; y) = \frac{1}{2}\mathrm{Tr}\left((s_1\widetilde{\Sigma}_n^{-1} yy^{\mathrm{T}}\widetilde{\Sigma}_n^{-1} - \widetilde{\Sigma}_n^{-1})\frac{\partial\widetilde{\Sigma}_n}{\partial\theta_k}\right) \tag{A13}$$

Let l be the length of θ and $s_T(\theta; y) = (s_1(\theta; y), \cdots, s_l(\theta; y))^{\mathrm{T}}$. The score function becomes that under the GPR model when $s_1 = 1$. The impact factor $s_1 = (n+2v)/(2(v-1)+y^{\mathrm{T}}\widetilde{\Sigma}_n^{-1}y)$ is very important for estimating θ. For example, when $y_j \to \infty$ for some j, the score function $s_T(\theta; y)$ is bounded, while that from the GPR model is not.

For a given parameter v, following Ref. [25] the influence function for the estimator $\widehat{\theta}$ is

$$\mathrm{IF}(y; \widehat{\theta}, F) = -\left(E\left(\frac{\partial^2 l(\theta; v)}{\partial\theta\partial\theta^{\mathrm{T}}}\right)\right)^{-1} s_T(\theta; y) \tag{A14}$$

Note that the matrix $\partial^2 l(\theta; v)/(\partial\theta\partial\theta^{\mathrm{T}})$ is bounded for y, which indicates that the influence function of $\widehat{\theta}$ is bounded under the meTPR model. Similarly, we can get that the score function is unbounded. So, for mGPR, the influence function of parameter estimation is also unbounded.

Proof of Proposition 4.2 Obviously $q^2 = (y-F_0(x))^{\mathrm{T}}(y-F_0(x))/\phi_0 = O(n)$. Under the condition of Lemma A.1 and the condition that $\| F_0 \|_k$ is bounded and $E_x(\log|I_n+\phi_0^{-1}\widetilde{K}_n|) = o(n)$ is established. According to Lemma A.1, for a positive constant c and any $\varepsilon>0$, when n is large enough, we have

$$\frac{1}{n}E_x(D[p_{\phi_0}(y \mid F_0, x), p_{\phi_0,\widehat{\theta}}(y \mid x)]) =$$

$$E_x\int\frac{1}{n}(-\log p_{\phi_0,\widehat{\theta}}(y \mid x) + \log p_{\phi_0}(y \mid F_0, x))\mathrm{d}p_{\phi_0,\widehat{\theta}}(y \mid x) \leqslant$$

$$E_x\int\left(\frac{1}{2n}\log|I_n + \phi_0^{-1}\widetilde{K}_n| + \frac{q^2+2(v-1)}{2n(n+2v-2)}(\|F_0\|_k^2 + c) + \frac{c}{n} + \varepsilon\right)\mathrm{d}p_{\phi_0,\widehat{\theta}}(y \mid x) \tag{A15}$$

It gives

$$\frac{1}{n}E_x(D[p_{\phi_0}(y \mid F_0, x), p_{\phi_0,\widehat{\theta}}(y \mid x)]) \to 0, \text{ as } n \to \infty \tag{A16}$$

Thus, the proposition holds.