Gaussian process methods for nonparametric functional regression with mixed predictors

Bo Wang*

Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK

Aiping Xu

Sigma (Maths and Stats Support), Coventry University, Coventry CV1 5DD, UK

Abstract

Gaussian process methods are proposed for nonparametric functional regression for both scalar and functional responses with mixed multidimensional functional and scalar predictors. The proposed models allow the response variables to depend on the entire trajectories of the functional predictors. They inherit the desirable properties of Gaussian process regression, and can naturally accommodate both scalar and functional variables as the predictors, as well as easy to obtain and express uncertainty in predictions. The numerical experiments show that the proposed methods significantly outperform the competing models, and their usefulness is also demonstrated by the application to two real datasets.

Keywords: Functional regression, Functional principal component analysis, Gaussian process regression, Nonparametric methods, Semi-metric

1. Introduction

The fast progress in information technology has provided the ability of recording very large datasets that offer the opportunity to observe phenomena in a more accurate way by generating data samples over very fine grids. The information in such high-dimensional datasets varies over some continuum which allows such data to be treated as a collection of mathematical objects, that is, as a collection of curves or of surfaces. For example, time

*Corresponding author. Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK. Tel.: +44-116-2522162. Fax: +44-116-2523915.

Email addresses: bo.wang@le.ac.uk (Bo Wang), aa9778@coventry.ac.uk (Aiping Xu)

series in financial engineering, imaging records in medicine, or spectrometric wavelengths in chemometrics can be considered as a discrete approximation of continuous variables. Such kind of data are termed as functional data, and the statistical methodology designated to deal with this kind of data is called functional data analysis (FDA). Under the functional data context, a curve is a random function, which is also considered as a sample from a stochastic process. FDA is the statistical analysis of datasets consisting of such random functions. FDA can reveal more statistical information contained in the smoothness and the derivatives of the functions, which distinguishes it from the multivariate data analysis (Ramsay and Silverman, 2005).

Among others one of the most important problems in FDA is functional regression which describes the relationship between the predictor and the response variable, where at least one of them is of functional nature. The first functional regression model was formulated by Hastie and Mallows (1993), and has ever since been extensively studied and further developed by Ramsay and Silverman (2005) and many other researchers.

For functional regression problems two main streams of methodologies exist in the literature: functional parametric models and functional nonparametric methods. Ramsay and Silverman (2005) studies in details the functional linear models and considers various cases such as scalar response with functional predictors, functional response with scalar predictors and functional response with functional predictors. Later on a large number of further extensions and developments have been proposed, for instance the generalised functional linear model (Müller and Stadtmüller, 2005), the Gaussian process functional regression models (Shi et al., 2007; Shi and Wang, 2008; Wang and Shi, 2014), the functional quadratic regression model (Yao and Müller, 2010), the penalized function-on-function regression (Ivanescu et al., 2015), and references therein. On the other hand, Ferraty and Vieu (2006) introduces functional nonparametric regression methods with functional predictors and scalar response, based on kernel-type methods. Baíllo and Grané (2009) proposes a local linear regression estimator and studies its asymptotic behaviour, and Ferraty et al. (2012) extends the kernel methods to the case of functional response. Preda (2007), Lian (2007) and Tang et al. (2015) use the reproducing kernel Hilbert spaces (RKHS) framework

in the nonparametric functional regression. Müller et al. (2013) and McLean et al. (2014) propose the continuously additive models and the functional generalized additive model for the scalar response case, respectively.

In this paper we introduce Gaussian process methods for nonparametric functional regression, where the response can be either scalar or functional with mixed multidimensional functional and scalar predictors. Gaussian process regression (GPR), as a nonparametric regression method, has been widely used and proven to be powerful and effective in various fields, due to many desirable properties, such as the existence of explicit forms, the ease of obtaining and expressing uncertainty in predictions, the ability to capture a wide variety of behaviour through covariance functions, and a natural Bayesian interpretation. We refer to the seminal book by Rasmussen and Williams (2006) for details. Shi et al. (2007) first applies GPR methods to functional data, which is further developed by Shi and Wang (2008), Shi and Choi (2011) and Wang and Shi (2014). However in the above works the functional response variable depends on the functional predictors at the current time only, therefore their models are types of the concurrent functional models (Ramsay and Silverman, 2005; Maity, 2017). In this paper we propose Gaussian process methods for functional regression where the response variable depends on the entire trajectories of the functional predictors. The proposed methods enjoy the intrinsic desirable properties of GPR, and can naturally incorporate both scalar and functional variables of high dimension as the predictors, as well as easy to obtain the predictive variance. Since bandwidth selection, which usually needs cross validation, is not required in our models, the proposed methods are much faster in computation than some of the existing kernel methods considered in the numerical examples at the same time of significantly improved prediction accuracy. And due to the nature of GPR, the dimension of the predictors has little impact on the computational time, so the models can easily deal with high dimensional predictors. Our numerical examples show that the proposed methods significantly outperform the existing methods in comparison.

The rest of the paper is organised as follows. Section 2 briefly reviews the GPR methods and then introduces the Gaussian process nonparametric regression methods for functional data with scalar response and functional response. The proposed methods are evaluated

by the simulation studies in Section 3, and are applied to two real datasets in Section 4. Finally, Section 5 concludes the paper with some discussions.

2. Methodology

2.1. Gaussian process regression

This section provides a brief introduction to Gaussian process regression methods; see Rasmussen and Williams (2006) for more complete discussion.

Consider a problem of nonlinear regression

$$y = f(x) + \varepsilon,$$

where the function $f(\cdot) : \mathbb{R}^p \mapsto \mathbb{R}$ is unknown and needs to be estimated. By Gaussian process method $f(\cdot)$ is assumed to follow a Gaussian process with mean function $\mu(\cdot)$ and covariance function $k(\cdot, \cdot)$. Given n pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$, we have

$$y_i = f(x_i) + \varepsilon_i,$$

where $\{\varepsilon_i\}_{i=1,\ldots,n}$ are independent and identically distributed normal random noises with mean 0 and variance σ^2. It follows that $(y_1, \ldots, y_n)^T$ has an n-variate normal distribution

$$(y_1, \ldots, y_n)^T \sim N(\boldsymbol{\mu}, K),$$

where $\boldsymbol{\mu} = (\mu(x_1), \ldots, \mu(x_n))^T$ is the mean vector and K is the $n \times n$ covariance matrix of which the (i, j)-th element $K_{ij} = k(x_i, x_j) + \sigma^2 \delta_{ij}$. Here $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

Let x^* be any test point and y^* be the corresponding response value. Then the joint distribution of $(y_1, \ldots, y_n, y^*)^T$ is an $(n + 1)$-variate normal distribution with mean $(\mu(x_1), \ldots, \mu(x_n), \mu(x^*))^T$ and covariance matrix

$$\begin{bmatrix} K & K^* \\ K^{*T} & k(x^*, x^*) + \sigma^2 \end{bmatrix}$$

where $K^* = (k(x^*, x_1), \cdots, k(x^*, x_n))^T$.

The conditional distribution of y^*, given $\mathbf{y} = (y_1, \cdots, y_n)^T$, is then $N(\hat{y}^*, \hat{\sigma}^{*2})$ with

$$\begin{aligned} \hat{y}^* &= \mu(x^*) + K^{*T} K^{-1} (\mathbf{y} - \boldsymbol{\mu}), \\ \hat{\sigma}^{*2} &= k(x^*, x^*) + \sigma^2 - K^{*T} K^{-1} K^*. \end{aligned}$$

In GPR, the covariance function $k(\cdot,\cdot)$ plays a crucial role in the predictive mean and variance. Covariance functions contain our presumptions about the function we wish to learn and define the closeness and similarity between data points. As a result, the choice of covariance function may have profound impacts on the performance of a GPR model. A wide range of covariance functions have been proposed and discussed; see for example Rasmussen and Williams (2006) and Shi and Choi (2011). In this paper we will adopt the most commonly used covariance function - the squared exponential covariance function:

$$k(x_i, x_j) = v \exp\left(-\frac{1}{2}\sum_{d=1}^{p} w_d (x_{id} - x_{jd})^2\right), \tag{1}$$

and assume the mean function $\mu(x)$ to be 0.

The hyper-parameters $\{v, w_1, \ldots, w_p\}$ in (1) and the noise variance σ^2 can be estimated by the maximum likelihood method. The log-likelihood of the training data is given by

$$L(v, w_1, \ldots, w_p, \sigma^2) = -\frac{1}{2}\log\det K - \frac{1}{2}\mathbf{y}^T K^{-1}\mathbf{y} - \frac{n}{2}\log 2\pi.$$

And the derivative of the log-likelihood with respect to each parameter (denoted by a generic notation θ) is:

$$\frac{\partial L}{\partial \theta} = -\frac{1}{2}\mathrm{tr}\left(K^{-1}\frac{\partial K}{\partial \theta}\right) + \frac{1}{2}\mathbf{y}^T K^{-1}\frac{\partial K}{\partial \theta}K^{-1}\mathbf{y}.$$

Hence standard gradient based numerical optimisation techniques, such as Conjugate Gradient method, can be used to maximise the log-likelihood function $L(v, w_1, \ldots, w_p, \sigma^2)$ to obtain the estimates of the parameters.

2.2. Gaussian process for functional regression

2.2.1. Functional regression with scalar response and mixed predictors

Now suppose that y is a scalar response in $\mathbb{R}$, $X(t)$ a q-dimensional functional predictor ($t \in \mathcal{T}$), and z a p-dimensional scalar predictor. Consider the problem of nonlinear regression

$$y = f(X, z) + \varepsilon,$$

where $f(\cdot,\cdot)$ is an unknown functional operator, and $\varepsilon \sim N(0,\sigma^2)$. We assume that $f(\cdot,\cdot)$ follows a Gaussian process with mean 0 and covariance function $k(\cdot,\cdot,\cdot,\cdot)$, defined as

$$k(X_i, X_j, z_i, z_j) = v \exp\left(-\frac{1}{2}\sum_{d=1}^{p} w_d (z_{id} - z_{jd})^2 - \frac{1}{2}\sum_{d=1}^{q} \eta_d \|X_{id} - X_{jd}\|_d^2\right), \qquad (2)$$

where $\|\cdot\|_d$ denotes the semi-metric defined for the dth component of the functional predictor. Semi-metrics as a closeness measure for functional data are discussed in Ferraty and Vieu (2006), and it is demonstrated that semi-metric spaces are better adapted than metric spaces in the functional context. Ferraty and Vieu (2006) introduces three families of semi-metrics, which are based on functional principal component analysis (FPCA), on derivatives and on partial least squares (PLS), respectively. The first two of them are used in our numerical examples so are briefly presented below for completeness.

Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be a sample of curves which are identically distributed as the functional random variable $\mathcal{X} = \{\mathcal{X}(t); t \in \mathcal{T}\}$.

Semi-metric based on FPCA is defined as

$$d_q^{PCA}(\mathcal{X}_i, \mathcal{X}_j) = \sqrt{\sum_{k=1}^{q} \left(\int [\mathcal{X}_i(t) - \mathcal{X}_j(t)] v_k(t) dt\right)^2},$$

where $v_1, \ldots, v_q$ are the orthonormal eigenfunctions of the covariance operator $\Gamma_{\mathcal{X}}(s,t) = E(\mathcal{X}(s)\mathcal{X}(t))$ associated with the largest q eigenvalues.

Semi-metric based on derivatives is defined as

$$d_q^{deriv}(\mathcal{X}_i, \mathcal{X}_j) = \sqrt{\int \left(\mathcal{X}_i^{(q)}(t) - \mathcal{X}_j^{(q)}(t)\right)^2 dt},$$

where $\mathcal{X}^{(q)}$ denotes the qth order derivative of $\mathcal{X}$.

We refer to Ferraty and Vieu (2006) for the practical implementation of the semi-metrics. As commented in the above book, FPCA-type semi-metrics are suitable for rough datasets, while derivatives-type ones are adapted to smooth datasets. This rule of thumb will be adopted in our numerical examples. How to choose the best semi-metric in practice remains an open question (Ferraty and Vieu, 2006). It should be noted that the semi-metrics in (2) can be chosen differently for different components of the functional predictor as appropriate.

Given n observations $(X_1, z_1, y_1), \ldots, (X_n, z_n, y_n)$, it yields that $(y_1, \ldots, y_n)^T$ has an n-variate normal distribution

$$\mathbf{y} = (y_1, \ldots, y_n)^T \sim N(\mathbf{0}, K),$$

where K is the $n \times n$ covariance matrix with the (i, j)-th element $K_{ij} = k(X_i, X_j, z_i, z_j) + \sigma^2 \delta_{ij}$. Hence the parameters in the model $\{v, w_1, \ldots, w_p, \eta_1, \ldots, \eta_q, \sigma^2\}$ can be estimated by maximising the following log-likelihood function

$$L(v, w_1, \ldots, w_p, \eta_1, \ldots, \eta_q, \sigma^2) = -\frac{1}{2} \log \det K - \frac{1}{2} \mathbf{y}^T K^{-1} \mathbf{y} - \frac{n}{2} \log 2\pi.$$

Let (X^*, z^*) be a test input and y^* be the corresponding response value. Following the same argument as in Subsection 2.1, the predictive mean $\hat{y}^*$ and variance $\hat{\sigma}^{*2}$ of y^* are given by

$$\begin{aligned} \hat{y}^* &= K^{*T} K^{-1} \mathbf{y}, \\ \hat{\sigma}^{*2} &= k(X^*, X^*, z^*, z^*) + \sigma^2 - K^{*T} K^{-1} K^*, \end{aligned}$$

where $K^* = (k(X^*, X_1, z^*, z_1), \cdots, k(X^*, X_n, z^*, z_n))^T$.

2.2.2. Functional regression with functional response and mixed predictors

Let $Y(t)$ be an L_2-continuous stochastic process on $\mathcal{T}$, $\mu(t)$ be its mean function and $C(t, t')$ its covariance function. By the functional principal component analysis (FPCA), $Y(t)$ can be decomposed as

$$Y(t) = \mu(t) + \sum_{j=1}^{\infty} \beta_j \phi_j(t), \quad t \in \mathcal{T},$$

where β_j's are uncorrelated random variables with mean zero and variance λ_j, $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues and $\phi_1(t), \phi_2(t), \ldots$ are the associated eigenfunctions of the covariance function $C(\cdot, \cdot)$.

Now we consider the following functional regression model with functional response

$$Y_i(t) = f(X_i(\cdot), z_i) + \varepsilon_i(t), \tag{3}$$

where $f(\cdot, \cdot)$ is an unknown functional operator, depending on a q-dimensional functional predictor $X(t)$ and a p-dimensional scalar predictor z, $\varepsilon_i(t)$ $(i = 1, \ldots, n)$ are independent

white noise processes with variance σ_ε^2, and $Y_i(t)$ $(i = 1, \ldots, n)$ are independent samples from $Y(t)$ with noises.

Let $\tilde{Y}_i(t)$ be the smoothed version of $Y_i(t)$, then by FPCA, we have

$$\tilde{Y}_i(t) = \mu(t) + \sum_{j=1}^{\infty} \beta_{ij}\phi_j(t), \quad t \in \mathcal{T}, \tag{4}$$

where, for any $j > 0$ and $i = 1, \ldots, n$, $\beta_{ij} = \int_{\mathcal{T}}(\tilde{Y}_i(t) - \mu(t))\phi_j(t)dt$ is the jth principal component score of the ith sample.

To obtain an approximation of $\tilde{Y}_i(t)$, we truncate (4) at the first J terms so that

$$\tilde{Y}_i(t) = \mu(t) + \sum_{j=1}^{J} \beta_{ij}\phi_j(t), \quad t \in \mathcal{T}. \tag{5}$$

Therefore the regression function $f(\cdot, \cdot)$ can be represented by the relationships between β_{ij} and $X_i(t)$ and z_i, that is, for any j $(j = 1, \ldots, J)$,

$$\beta_{ij} = r_j(X_i(t), z_i) + e_{ij}, \quad i = 1, \ldots, n,$$

where $e_{ij} \sim N(0, \sigma_j^2)$ and $r_j(\cdot, \cdot)$ is a functional operator representing the relationship between the jth principal component of $Y(t)$ and the predictors. In this paper we propose to use Gaussian process methods to model $r_j(\cdot, \cdot)$, that is, for $j = 1, \ldots, J$, $r_j(\cdot, \cdot)$ is assumed to follow a Gaussian process with mean function 0 and the following covariance function

$$k^{(j)}(X_l, X_m, z_l, z_m) = v^{(j)} \exp\left(-\frac{1}{2}\sum_{d=1}^{p} w_d^{(j)}(z_{ld} - z_{md})^2 - \frac{1}{2}\sum_{d=1}^{q} \eta_d^{(j)} \|X_{ld} - X_{md}\|_d^2\right).$$

Thus, for a given j, we have n training data points $(X_1, z_1, \beta_{1j}), \ldots, (X_n, z_n, \beta_{nj})$ in the Gaussian process regression model, where $\{\beta_{ij}\}_{i=1}^n$ are the response values and $\{X_i\}_{i=1}^n$ and $\{z_i\}_{i=1}^n$ are the corresponding functional and scalar predictor values. It follows that $(\beta_{1j}, \ldots, \beta_{nj})^T$ has an n-variate normal distribution

$$\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{nj})^T \sim N(\mathbf{0}, K^{(j)}),$$

where $K^{(j)}$ is the $n \times n$ covariance matrix with the (l, m)-th element $K_{lm}^{(j)} = k^{(j)}(X_l, X_m, z_l, z_m) + \sigma_j^2\delta_{lm}$. Therefore the parameters $\{v^{(j)}, w_1^{(j)}, \ldots, w_p^{(j)}, \eta_1^{(j)}, \ldots, \eta_q^{(j)}, \sigma_j^2\}$ can be estimated by maximising the following log-likelihood function

$$L(v^{(j)}, w_1^{(j)}, \ldots, w_p^{(j)}, \eta_1^{(j)}, \ldots, \eta_q^{(j)}, \sigma_j^2) = -\frac{1}{2}\log\det K^{(j)} - \frac{1}{2}\boldsymbol{\beta}_j^T(K^{(j)})^{-1}\boldsymbol{\beta}_j - \frac{n}{2}\log 2\pi.$$

Hence, given a test input $(X^*(t), z^*)$, let $Y^*(t)$ be the corresponding functional response and β_j^* be its jth principal component score for $j = 1, \ldots, J$. Then the predictive mean $\hat{\beta}_j^*$ and variance $\hat{\sigma}_j^{*2}$ of β_j^* can be obtained by

$$\begin{aligned} \hat{\beta}_j^* &= (K^{(j)*})^T (K^{(j)})^{-1} \boldsymbol{\beta}_j, \\ \hat{\sigma}_j^{*2} &= k^{(j)}(X^*, X^*, z^*, z^*) + \sigma_j^2 - (K^{(j)*})^T (K^{(j)})^{-1} K^{(j)*}, \end{aligned}$$

where $K^{(j)*} = (k^{(j)}(X^*, X_1, z^*, z_1), \cdots, k^{(j)}(X^*, X_n, z^*, z_n))^T$.

Consequently the predictive mean and variance of the functional response $Y^*(t)$ are given by

$$\hat{Y}^*(t) = \hat{\mu}(t) + \sum_{j=1}^{J} \hat{\beta}_j^* \phi_j(t), \tag{6}$$

$$\hat{\sigma}^{*2}(t) = \hat{\sigma}_\mu^2(t) + \sum_{j=1}^{J} \hat{\sigma}_j^{*2} \phi_j^2(t) + \hat{\sigma}_\varepsilon^2, \tag{7}$$

where $\hat{\sigma}_\varepsilon^2$ denotes the variance of $\varepsilon(t)$ which can be estimated from the smoothing method used (Ramsay and Silverman, 2005), and $\hat{\mu}(t)$ and $\hat{\sigma}_\mu^2(t)$ are the estimated mean function and its variance respectively which can be obtained by, for example, local linear estimator (Degras, 2011) or polynomial spline estimator (Cao et al., 2012).

In practice, the functional response and the functional predictors are observed at discrete observation points, so the eigenfunctions and the principal component scores can be obtained numerically (Ramsay and Silverman, 2005). Besides, the number of principal components J plays an important role on the accuracy of prediction. In practice J can be chosen by cross validation, or such that (5) provides a good approximation, for example, such that the cumulative percentage of the variation explained by the first J components, measured by the ratio $\sum_{j=1}^{J} \lambda_j / \sum_{j=1}^{\infty} \lambda_j$, exceeds 99% of the total variation.

3. Simulation studies

To demonstrate the effectiveness of the GPR methods we conduct two simulation examples in this section: one is for the case with scalar response and functional predictor and the other for functional response and functional predictor. Because not all the existing models

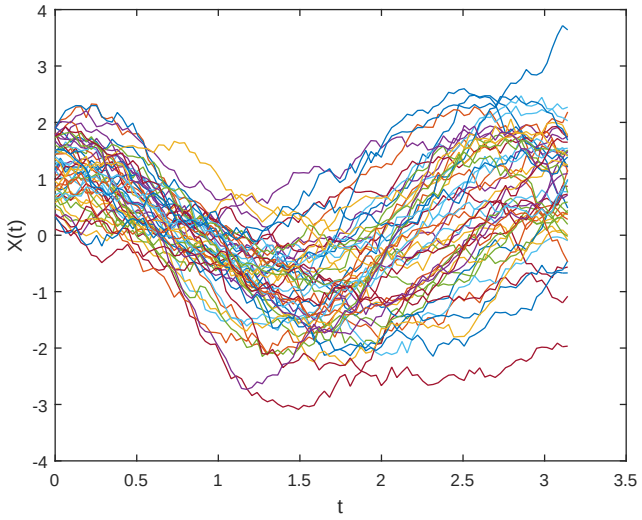


Figure 1: Sample predictor curves for the case of scalar response.

used for comparison are able to deal with mixed predictors, only functional predictors are considered in the simulation studies.

Scalar response. The data are simulated as follows. Let $X_1, \ldots, X_n$ be $n = 50$ samples of a functional predictor such that

$$X_i(t_j) = a_i \cos(\omega_i t_j) + \sum_{k=1}^{j} W_{ik}, \tag{8}$$

where $a_1, \ldots, a_n$ are n independent real random variables uniformly distributed in $[0.2, 2]$, $\omega_1, \ldots, \omega_n$ are n independent real random variables uniformly distributed in $[1.5, 2.5]$, $0 = t_1 < t_2 < \cdots < t_{100} = \pi$ are equally spaced points, and the W_{ik}'s are i.i.d. samples of a normal distribution with mean 0 and variance 0.01.

The regression function $f(\cdot)$ and the response variable are defined as, for $i = 1, \ldots, n$,

$$f(X_i) = \int_0^\pi X_i^2(t)dt, \quad y_i = f(X_i) + \varepsilon_i$$

with $\varepsilon_i \sim N(0, 0.04)$. An example of the predictor curves is shown in Figure 1.

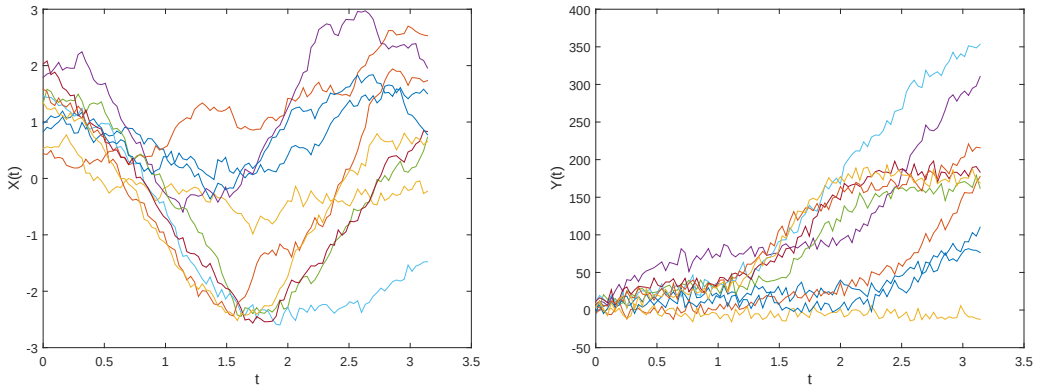
As the predictor curves are not smooth, the FPCA-type semi-metric is used in our model. The performance of the proposed Gaussian process nonparametric regression method (GPNR) is compared with three existing nonparametric methods and two frequently used

Table 1: Means and standard deviations (in brackets) of squared prediction errors for scalar response

Kernel	LL-Trig	LL-Eig	FL-Eig	FL-Trig	GPNR
0.1681	0.0698	0.0680	0.4926	0.4813	0.0259
(0.5003)	(0.1339)	(0.1311)	(1.1121)	(1.1482)	(0.0759)

parametric functional linear models, namely: the kernel estimator discussed in Ferraty and Vieu (2006) (Kernel), the local linear estimator proposed by Baíllo and Grané (2009) with the Fourier trigonometric basis (LL-Trig) and with the eigenfunctions (LL-Eig), the functional linear models based on the trigonometric expansion (FL-Trig) and on the eigenfunctions (FL-Eig) (Ramsay and Silverman, 2005; Cai and Hall, 2006). For both the kernel and the local linear estimators the Gaussian kernel is used, and the kernel bandwidth and the number of terms in the series expansion are chosen via cross validation procedure; see Baíllo and Grané (2009) for more details on the implementation of the estimators. FL-Trig is the linear estimator based on the trigonometric expansion of the covariates and regularised using roughness penalties, while FL-Eig is based on expanding the functional predictor in terms of its covariance eigenfunctions. The cut-off in the expansion and the smoothing parameter in the penalties are chosen via cross validation. The means and the standard deviations of the squared prediction errors between the true regression values and the predictions obtained from the above six methods via leave-one-out cross validation are calculated. The above procedure is repeated for 20 times and the averages of the means and the standard deviations are reported in Table 1.

Since the regression function is highly nonlinear, the functional linear models perform much worse than the nonparametric methods as expected. And, among the latter, the GPNR significantly outperforms the others, and our experiment shows that it is much faster than the local linear estimators (2.3 seconds by GPNR versus 44.8 seconds by LL-Trig and 236.8 seconds by LL-Eig on our desktop computer for each repetition) and is comparable with the kernel estimator (2.4 seconds).



(a) Functional predictor (b) Functional response

Figure 2: Sample predictor curves and response curves for the case of functional response.

Functional response. For the ease of comparison, we adopt the same simulation example as in Ferraty et al. (2012), which is described as follows for completeness.

Let $X_1, \ldots, X_n$ be $n = 250$ samples of a functional predictor defined by (8). The regression function $f(\cdot)$ and the response variable are defined as, for $i = 1, \ldots, 250$ and $j = 1, \ldots, 100$,

$$f(X_i)(t_j) = \int_0^{t_j} X_i^2(u)du, \quad Y_i(t_j) = f(X_i)(t_j) + \varepsilon_i(t_j),$$

where the error term $\varepsilon_i(t_j)$ is the mixture of the additive and structural errors, that is, $\varepsilon_i^{\text{mix}}(t_j)$ described in Ferraty et al. (2012). As demonstration ten predictor curves and the corresponding response curves are shown in Figure 2.

We split the original sample into two sets: the learning sample $(X_i, Y_i)_{i=1,\ldots,200}$ and the test sample $(X_i, Y_i)_{i=201,\ldots,250}$. In the Gaussian process nonparametric regression method (GPNR), the B-spline basis is used to smooth the response curves $Y_i(t)$, and the FPCA decomposition is then applied to the smoothed curves. The FPCA-type semi-metric is adopted for the functional predictors as the curves are not smooth. To examine the effect of the number of principal components (PCs) on the accuracy of the prediction, we use different choices in the model and calculate the mean squared prediction errors between the predicted and the true regression values for the fifty test samples. They are presented

Table 2: Mean squared prediction errors (MSE) for functional response case with different number of principal components (PCs)

Number of PCs	2	3	4	5	6	7
Cumulative %	95.79	98.89	99.62	99.83	99.93	99.96
MSE	114.8792	43.0786	23.7840	22.1508	19.6952	19.6952

in Table 2.

We also compare the performance of our model with four existing methods: the kernel method discussed in Ferraty et al. (2012) (Kernel), the functional additive model proposed by Müller and Yao (2008) (FAM), the penalized function-on-function regression proposed by Ivanescu et al. (2015) (PFFR), and the functional linear model (FLM) studied in Yao et al. (2005). The kernel method is performed using R routines available at https://www.math.univ-toulouse.fr/staph/npfda/, and the detailed implementation of this method is provided in Ferraty et al. (2012). FAM and FLM are conducted using PACE package (http://www.stat.ucdavis.edu/PACE/), where Gaussian kernel is used to estimate the additive model components and generalised cross-validation (GCV) is used to choose the tuning parameters. PFFR is implemented using the function `pffr()` in the R package 'refund'. The mean squared prediction errors by these four methods are presented in Table 3.

It can be seen from Table 2 that the prediction accuracy increases with the increasing number of PCs in the GPNR model until the number of PCs reaches six when the cumulative percentage of the variation accounted for by the selected PCs is nearly 100% of the total variation and the prediction accuracy can not be improved any further. And it is apparent that our GPNR model significantly outperforms the other four models, even using only two PCs in the model, and the Kernel method performs the second best. Since the regression function is nonlinear, the PFFR which deals with linear function-on-function regression and the FLM do not provide satisfactory predictions as can be expected.

The prediction results by the GPNR with six PCs, Kernel and FAM for four randomly selected samples are shown in Figure 3.

Table 3: Mean squared prediction errors for functional response

Kernel	FAM	PFFR	FLM
377.1323	450.0289	1552.72	1704.25

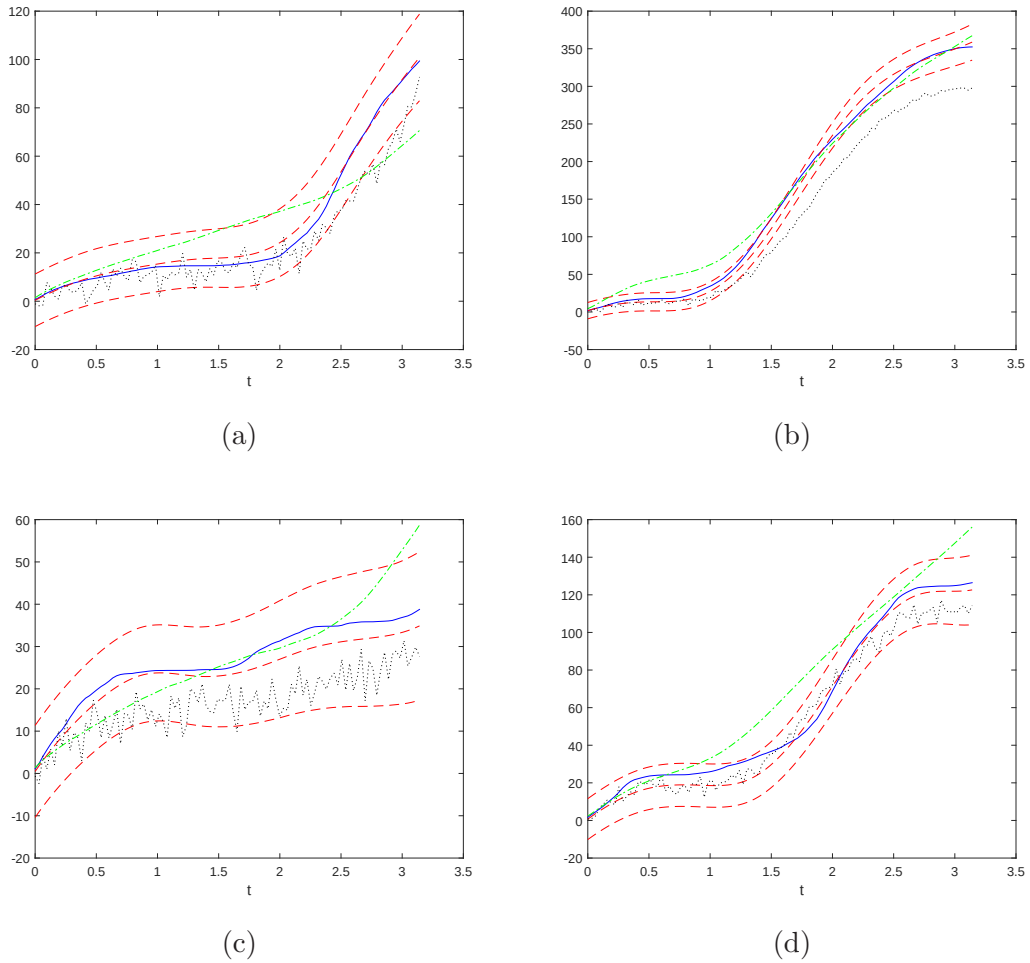


Figure 3: Four randomly selected predictions for functional response. Solid lines are the true regression curves, dashed lines are the prediction by GPNR (with 95% confidence bands), dotted lines by Kernel, and dash-dot lines by FAM.

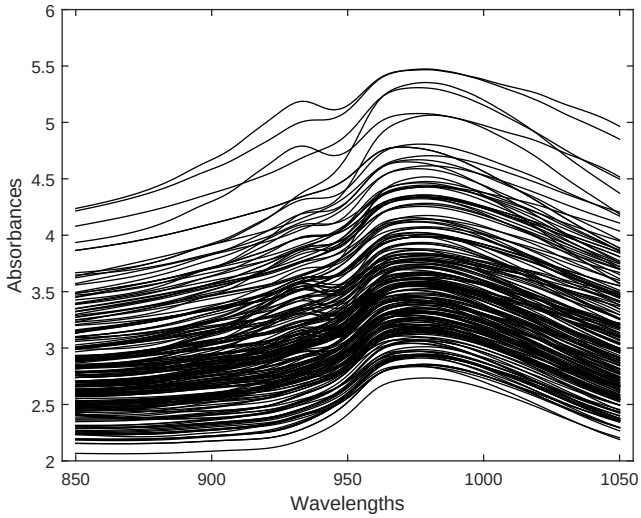


Figure 4: A graphical display of spectrometric curves.

4. Real data

We now apply our Gaussian process nonparametric regression methods to two real datasets.

4.1. Spectrometric data

The first dataset is the spectrometric data, which have been studied by Ferraty and Vieu (2006) and are downloadable from the *Nonparametric Functional Data Analysis* website (http://www.math.univ-toulouse.fr/staph/npfda/). We refer to Ferraty and Vieu (2006) for the detailed description of the data. The spectrometric curves are shown in Figure 4. The task is to predict fat content from the spectrometric curves, therefore the response is the fat content and the predictor is the spectrometric curves.

As done in Ferraty and Vieu (2006), the original sample is split into two subsamples: the learning sample contains the first 160 units and the testing sample contains the last 55 units. Since the spectrometric curves are smooth, the semi-metric based on derivative of order two is used in our model. The mean squared prediction errors by various methods are reported in Table 4.

It is noted that a slightly different spectrometric dataset was studied by Yao and Müller

Table 4: Mean squared prediction errors for spectrometric data

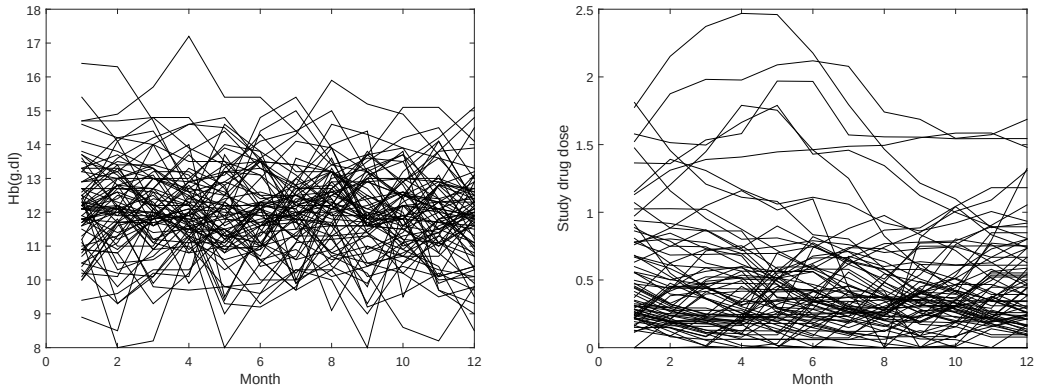
LL-Trig	LL-Eig	FL-Eig	FL-Trig	GPNR
4.4495	3.7255	7.5261	6.8888	2.1157

(2010) to demonstrate the usefulness of their proposed functional quadratic regression model (FQR). When applied to the above data, the mean squared prediction error by FQR is 148.0210. Also recall that in Section 7.2 of Ferraty and Vieu (2006) the mean squared prediction errors for the same dataset using the kernel methods are as follows: 3.5 by the conditional expectation (i.e. regression) method, 3.61 by the conditional mode method, and 3.44 by the conditional median method. It is obvious that the prediction accuracy by our model is significantly better than all the competitors.

4.2. Leeds renal anaemia data

Patients with reduced kidney function not only require dialysis to remove waste products from their blood, but they also produce less erythropoietin (EPO), the natural stimulus to the production of red blood cells in the bone marrow. As a consequence most dialysis patients suffer renal anaemia to some degree and require regular injections with either a synthetic EPO, for example Erythropoietin Beta (EB), or a modified epoetin such as Darbepoetin Alpha (DA). The dose of epoetin to be given to each patient is determined by monitoring the haemoglobin (Hb) concentration from a blood sample, such that their Hb levels are controlled within relatively narrow limits. If Hb levels are too low then patients become symptomatic of anaemia and if too high then there may be pro-thrombotic risks to their dialysis treatment and vascular tree. The primary therapeutic concern is how to maintain the Hb level of each patient by giving a suitable dose of exogenous epoetin.

In this example, we look at the data collected from 74 patients who received DA in Leeds, UK. Figure 5 shows the monthly Hb measurements recorded and the dose levels of the agent DA received for these patients for the period of 12 months. More detailed description of the data is given in Shi et al. (2012). The objective is to use the dose levels to predict the Hb levels, hence the functional response is the Hb level and the dose level of



(a) (b)

Figure 5: Leeds renal data: (a) the Hb levels recorded for 74 patients for the period of 12 months; (b) the dose levels of the agent DA received by the patients in the same period.

Table 5: Mean squared prediction errors (MSE) for renal anaemia data with different number of principal components (PCs)

Number of PCs	2	3	4	5	6	7	8	9
Cumulative %	99.3	99.5	99.7	99.8	99.9	99.95	99.97	100.00
MSE	1.5778	1.3528	1.1977	1.1657	1.1182	1.1267	1.1242	1.1237

DA is the functional predictor.

We apply the Gaussian process nonparametric regression method (GPNR) to the data, where the B-spline basis is used to smooth the response curves and the FPCA-type semimetric is adopted for the functional predictor. The mean squared prediction errors by leave-one-out cross validation using different number of principal components in the model are presented in Table 5 and plotted in Figure 6. For comparison the four existing models (Kernel, FAM, PFFR and FLM) are also applied to the same data, and the mean squared prediction errors by these four methods are given in Table 6.

As shown in Table 5 and Figure 6, the prediction accuracy increases with the increasing number of PCs in the GPNR model until the number of PCs reaches six when the

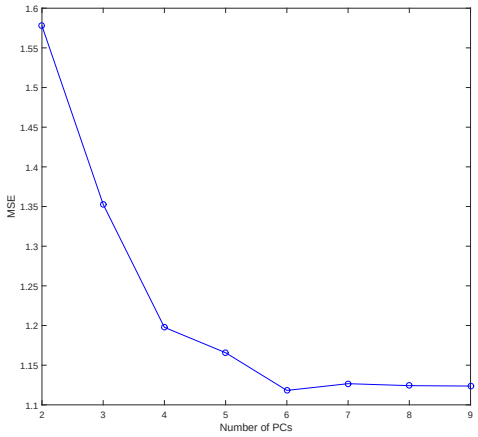


Figure 6: Leeds renal data: MSE for different number of principal components.

Table 6: Mean squared prediction errors for renal anaemia data with different models

Kernel	FAM	PFFR	FLM
1.2861	1.9622	1.1303	1.9479

cumulative percentage of the variation accounted for by the selected PCs is 99.9% of the total variation, after which the MSE has slight increase and remains almost constant. The proposed GPNR model with six PCs provides the most accurate predictions among all the models in comparison. Even using only four PCs, the GPNR model still outperforms the Kernel, FAM and FLM methods, and just slightly worse than the PFFR model.

5. Conclusion

We have introduced Gaussian process methods for nonparametric functional regression with either scalar response or functional response. Unlike the Gaussian process functional regression models proposed in Shi et al. (2007) which is a type of the concurrent functional models, our proposed methods allow response variables to depend on the entire trajectories of the functional predictors. The proposed methods provide a flexible yet efficient framework for nonparametric functional regression. They inherit the desirable properties of GPR methods, and are able to incorporate prior knowledge and specifications about the regression function and the proximity between functional data by selecting different covariance functions. The proposed methods naturally incorporate both multiple scalar and multiple functional variables as the predictors, which has not been studied in the literature in the context of nonparametric functional regression, and the predictive variance (hence the uncertainty in prediction) can easily be obtained. The numerical experiments show that the proposed methods significantly outperform the competing methods, and in the case of scalar response it is much faster than the closest competitor (local linear regression methods). Another advantage of the proposed methods is that, since the dimension of the predictors has little impact on the computational time, our methods are able to cope with very high dimensional scalar and functional predictors.

In this paper we only consider the most commonly used covariance function - squared exponential covariance function. Although the numerical examples show it is effective and very robust, it will be useful to explore other covariance functions in this context, such as Matérn class (Rasmussen and Williams, 2006) and the spectral mixture (Wilson and Adams, 2013), and the sensitivity of the methods to the type of covariance functions.

Another possible extension of the work concerns nonparametric methods for functional data clustering. Ferraty and Vieu (2006) studies this problem using kernel-type methods. It will be interesting to investigate how the Gaussian process nonparametric methods perform in this aspect.

Acknowledgement

The authors thank the Associate Editor and the reviewers for their constructive suggestions and very helpful comments.

References

Baíllo, A., Grané, A., 2009. Local linear regression for functional predictor and scalar response. Journal of Multivariate Analysis 100 (1), 102–111.

Cai, T. T., Hall, P., 2006. Prediction in functional linear regression. Ann. Statist. 34 (5), 2159–2179.

Cao, G., Yang, L., Todem, D., 2012. Simultaneous inference for the mean function based on dense functional data. J. Nonparametr. Stat. 24 (2), 359–377.

Degras, D. A., 2011. Simultaneous confidence bands for nonparametric regression with functional data. Statist. Sinica 21 (4), 1735–1765.

Ferraty, F., Van Keilegom, I., Vieu, P., 2012. Regression when both response and predictor are functions. Journal of Multivariate Analysis 109, 10–28.

Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer, New York.

Hastie, T., Mallows, C., 1993. A statistical view of some chemometrics regression tools: Discussion. Technometrics 35 (2), 140–143.

Ivanescu, A. E., Staicu, A.-M., Scheipl, F., Greven, S., 2015. Penalized function-on-function regression. Computational Statistics 30 (2), 539–568.

Lian, H., 2007. Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. Canad. J. Statist. 35 (4), 597–606.

Maity, A., 2017. Nonparametric functional concurrent regression models. Wiley Interdisciplinary Reviews: Computational Statistics 9 (2), e1394–n/a.

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., Ruppert, D., 2014. Functional generalized additive models. Journal of Computational and Graphical Statistics 23 (1), 249–269.

Müller, H.-G., Stadtmüller, U., 2005. Generalized functional linear models. Ann. Statist. 33 (2), 774–805.

Müller, H.-G., Wu, Y., Yao, F., 2013. Continuously additive models for nonlinear functional regression. Biometrika 100 (3), 607–622.

Müller, H.-G., Yao, F., 2008. Functional additive models. J. Amer. Statist. Assoc. 103 (484), 1534–1544.

Preda, C., 2007. Regression models for functional data by reproducing kernel Hilbert spaces methods. J. Statist. Plann. Inference 137 (3), 829–840.

Ramsay, J. O., Silverman, B. W., 2005. Functional Data Analysis. Springer, New York.

Rasmussen, C. E., Williams, C. K. I., 2006. Gaussian Processes For Machine Learning. The MIT Press.

Shi, J., Wang, B., Will, E., West, R., 2012. Mixed-effects Gaussian process functional regression models with application to dose–response curve prediction. Statistics in Medicine 31 (26), 3165–3177.

Shi, J. Q., Choi, T., 2011. Gaussian Process Regression Analysis For Functional Data. CRC Press, Boca Raton, FL.

Shi, J. Q., Wang, B., 2008. Curve prediction and clustering with mixtures of Gaussian process functional regression models. Statistics and Computing 18 (3), 267–283.

Shi, J. Q., Wang, B., Murray-Smith, R., Titterington, D. M., 2007. Gaussian process functional regression modeling for batch data. Biometrics 63 (3), 714–723.

Tang, X., Hong, Z., Hu, Y., Lian, H., 2015. Gaussian process models for nonparametric functional regression with functional responses. Comm. Statist. Theory Methods 44 (16), 3428–3445.

Wang, B., Shi, J. Q., 2014. Generalized Gaussian process regression model for non-Gaussian functional data. J. Amer. Statist. Assoc. 109 (507), 1123–1133.

Wilson, A., Adams, R., 2013. Gaussian process kernels for pattern discovery and extrapolation. In: Proceedings of The 30th International Conference on Machine Learning. pp. 1067–1075.

Yao, F., Müller, H.-G., 2010. Functional quadratic regression. Biometrika 97 (1), 49–64.

Yao, F., Müller, H.-G., Wang, J.-L., 2005. Functional linear regression analysis for longitudinal data. The Annals of Statistics 33 (6), 2873–2903.