Statistica Sinica **31** (2021), 53-78
doi:https://doi.org/10.5705/ss.202018.0296

NONPARAMETRIC RANDOM EFFECTS FUNCTIONAL REGRESSION MODEL USING GAUSSIAN PROCESS PRIORS

Zhanfeng Wang[1], Hao Ding[1], Zimu Chen[1] and Jian Qing Shi[2]

[1] *University of Science and Technology of China and* [2] *Newcastle University*

Abstract: For functional regression models with functional responses, we propose a nonparametric random-effects model using Gaussian process priors. The proposed model captures the heterogeneity nonlinearly and the covariance structure nonparametrically, enabling longitudinal studies of functional data. The model also has a flexible form of mean structure. We develop a procedure to estimate the unknown parameters and calculate the random effects nonparametrically. The procedure uses a penalized least squares regression and a maximum a posterior estimate, yielding a more accurate prediction. The statistical theory is discussed, including information consistency. Simulation studies and two real-data examples show that the proposed method performs well.

Key words and phrases: Functional linear model, function-on-function regression model, Gaussian process priors, nonlinear random effects.

1. Introduction

In modern data analysis, measurements of the studied subject are frequently recorded and stored as a curve or a surface with high frequency, making functional data analysis increasingly important. For a functional regression model with a functional response variable (output), a concurrent model assumes that the response at a certain point (which could be temporal or spatial) depends on functional covariates at the same point (e.g., see Wang and Shi (2014)). However, in practice, the response at a point often depends on part of or the whole curve or surface of the functional covariates. In our example presented later, data are collected from 38 assessed movements (functional covariates), and used to evaluate upper-limb function after stroke (response, measured by CAHAI, a clinical score indicating the impairment level of upper limbs). This is part of a home-based rehabilitation system that uses a set of video games called Circus

Corresponding author: Jian Qing Shi, School of Mathematics and Statistics, Newcastle University, Newcastle, UK. E-mail: j.q.shi@ncl.ac.uk.

Challenge to improve upper-limb function of stroke survivors (Serradilla et al. (2014); Shi et al. (2013)). Data for each movement (i.e., forward circle, sawing, orientation) were measured at a frequency of 60. These data are essential to measuring upper-limb function (Cheng, Shi and Eyre (2017)). Let $y_m(t)$ be the overall recovery level of the upper-limb function (CAHAI) at time t after a stroke for the mth patient. At time t, the patient played an assessment game, and functional data were collected from the movements of each hand, and represented as a complete curve of movements, $\boldsymbol{x}(s,t)$, where t denotes longitudinal time. Based on the discussion in Cheng, Shi and Eyre (2017), $\boldsymbol{x}(s,\cdot)$ and other covariates provide information that can be used to evaluate upper-limb function. However, we must also consider the longitudinal effect for each patient over time t. This motivates us to propose the following model:

$$y_m(t) = \boldsymbol{z}_m^\top(t)\boldsymbol{\nu} + \int_{S_t} \boldsymbol{x}_m^\top(s,t)\boldsymbol{\beta}(s,t)ds + \tau_m(\boldsymbol{z}_m(t), \boldsymbol{x}_m(\cdot,t)) + \epsilon_m(t), \qquad (1.1)$$

for $m = 1,\dots,M$, where $\boldsymbol{z}_m(t) = (z_{m1}(t),\dots,z_{mp}(t))^\top$ is a p-dimensional vector of covariates (e.g., the time after stroke and kinematic variables between two hands at time t). The model reasonably assumes that the concurrent relationship between the response and $\boldsymbol{z}_m(t)$ has a constant coefficient $\boldsymbol{\nu}$. However, the response depends on the whole curve of each covariate in $\boldsymbol{x}_m(s,t) = (x_{m1}(s,t),\dots,x_{mq}(s,t))^\top$, a q-dimensional vector of functional covariates. The q-dimensional functional coefficient $\boldsymbol{\beta}(s,t)$ measures how to attract information from the whole curve $\boldsymbol{x}(s,\cdot)$ via direction of s in $\boldsymbol{\beta}(s,\cdot)$, and that the model changes longitudinally along t. S_t is a prespecified region around point t, and $\epsilon_m(t)$ is an error function. For the movement data, S_t is a fixed interval, independent of the visiting time t.

To model the heterogeneity among patients (a severe problem in many cases, including movement data), we introduce nonlinear random effects τ_m, that depend on both the scalar covariates $\boldsymbol{z}_m(t)$ and the functional covariates $\boldsymbol{x}_m(\cdot,t)$. This differs from most existed methods, which depend on the scalar variables only. Gaussian process (GP) priors are used to model the random effects nonparametrically. The conventional GP regression model uses a GP prior based on a measure defined in a Euclidean space (Shi and Choi (2011); Shi et al. (2007)). Here, we first extend the definition to a functional space, and then use it to model nonlinear random effects that depend on mixed functional and scalar covariates.

Model (1.1) has a flexible form. When $\tau_m = 0$ and $\boldsymbol{x}_m(s,t) = \tilde{\boldsymbol{x}}_m(s)$, independent of t, model (1.1) becomes the conventional function-on-function linear model discussed in Ramsay and Silverman (2005). Owing to its complexity, the function-on-function model has received relatively little attention in the literature. The linear relationships between a functional response and predictors were first studied by Ramsay and Dalzell (1991). They considered the following model:

$$y_m(t) = \int_a^b \tilde{\boldsymbol{x}}_m^\top(s)\boldsymbol{\beta}(s,t)ds + \epsilon_m(t), \tag{1.2}$$

where a and b are prespecified constants. When b depends on time t, model (1.2) can be rewritten as

$$y_m(t) = \int_a^t \tilde{\boldsymbol{x}}_m^\top(s)\boldsymbol{\beta}(s,t)ds + \epsilon_m(t), \tag{1.3}$$

which is often called the historical functional regression model or retrospective functional regression model; see Malfait and Ramsay (2003) and Gervini (2015). It is also possible to restrict the functional historical effect to a certain lag in the past; for example Kim, Sent$\ddot{u}$rk and Li (2011) consider the integral interval for s as $[t-\delta_1, t-\delta_2]$, where $0 < \delta_2 < \delta_1 < T$, and $t \in [\delta_1, T]$. A special case is a zero lag ($\delta_1 = \delta_2 = 0$), resulting in a concurrent functional linear model (Ramsay and Silverman (2005)) or a varying-coefficient model (Hastie and Tibshirani (1993)).

Model (1.2) is well studied; see, for example, Yao, Müller and Wang (2005a,b), Müller and Yao (2008), Yuan and Cai (2010), Crambes and Mas (2013), Scheipl, Staicu and Greven (2015), Meyeret al. (2015), Sun et al. (2018), Kim et al. (2018), Luo and Qi (2017), and the references therein. Yao, Müller and Wang (2005b) used a functional principal component analysis (FPCA) to study the asymptotic properties of the parameters. Sun et al. (2018) applied a penalized least squares method based on a reproducing kernel Hilbert space to estimate $\boldsymbol{\beta}(s,t)$. Meyeret al. (2015) proposed a Bayesian approach.

In addition to the common mean structure, researchers has begun focusing on models' covariance structure, which can use personal characteristics to improve an inference, especially prediction. The importance of using random effects to improve prediction is discussed in Rao (2003) and Robinson (1991). Owing to the complexity of the model, few studies have examined the nonlinear random effects or covariance structure in the function-on-function regression model in (1.1). We propose a nonparametric random-effects model based on GP priors , and allow it to depend on both functional and scalar covariates. We develop a flexible and accurate procedure to estimate the unknown parameters and calculate the random effects using a penalized least squares regression and a maximum a posteriori estimate (MAP). The proposed method offers several advantages. First, it provides estimates of the functional regression coefficients, and thus, a common mean structure for all subjects with a longitudinal tuning setting. Second, it allows for nonlinear random effects, modeled by GP priors nonparametrically, thus, it allows both scalar and functional covariates. This addresses the problem of heterogeneity among subjects, and provides a more accurate prediction by using subject-specific data or characteristics other than the common mean structure. Third, we propose a novel method for constructing a kernel function for a GP to deal with the complex covariance structure. The model in (1.1) is therefore called a nonparametric random-effects functional regression model using GP priors. We also prove statistical properties including information consistency. Further discussion about process regression analysis can be found in, for example, Rasmussen and Williams (2006), Shi and Choi (2011), Wang and Shi (2014), Wang, Shi and Lee (2017).

The remainder of the paper is organized as follows. Section 2 describes how to define the random-effects model using GP priors, and how to derive its prediction distribution. The details of the estimation procedure are also provided in this section. The asymptotic properties are discussed in Section 3. Numerical studies and two real examples of Canadian weather data and movement data for stoke patients are given in Section 4. All proofs are presented in the Appendix.

2. Functional Regression Model Using GP Priors

2.1. Nonparametric random-effects functional regression

Let $\boldsymbol{u}_m(t) = (\boldsymbol{z}_m^\top(t), \boldsymbol{x}_m^\top(\cdot, t))^\top$ be covariates at time t, and let the observation data $\{(y_{mi} = y_m(t_i), \boldsymbol{u}_m(t_i)) : i = 1, \ldots, n, m = 1, \ldots, M\}$ satisfy model (1.1), or specifically, the following model (the true model):

$$y_{mi} = \boldsymbol{z}_m^\top(t_i)\boldsymbol{\nu}_0 + \int_{S_t} \boldsymbol{x}_m^\top(s, t_i)\boldsymbol{\beta}_0(s, t_i)ds + \tau_{0m}(\boldsymbol{z}_m(t_i), \boldsymbol{x}_m(\cdot, t_i)) + \epsilon_{mi}, \quad (2.1)$$

where t_i is an observed time. In addition, $\boldsymbol{\nu}_0, \boldsymbol{\beta}_0$, and τ_{0m} are the true values of $\boldsymbol{\nu}, \boldsymbol{\beta}$, and τ_m, respectively, and $\epsilon_{mi} = \epsilon_m(t_i)$ is an error term. To fit model (2.1), we apply model (1.1) to estimate $\boldsymbol{\nu}_0$ and $\boldsymbol{\beta}_0$, and to predict τ_{0m}. In (1.1), if we focus on certain points over time t, then $y_m(t)$ can be treated as a longitudinal response. In general, for $t \in [a, b]$, $y_m(t)$ is a functional response variable. Note that although we treat $y_m(t)$ as a functional variable, our results are all readily extendable to longitudinal data.

Suppose that functional responses $\{y_1(\cdot), \ldots, y_M(\cdot)\}$ are independent, and that $\boldsymbol{\beta}(s,t)$ is a smooth and square integrable function. Let $GP(\mu, k)$ denote a GP with a mean function μ and a covariance kernel function k. The residual function satisfies $\epsilon_m \sim GP(0, \sigma^2\delta_\epsilon)$, where $\sigma^2 > 0$, $\delta_\epsilon(t_i, t_j) = I(t_i = t_j)$, and $I(\cdot)$ is an indicator function. The random effect τ_m has a GP prior with mean zero and covariance kernel function $k(\cdot, \cdot)$. Shi and Choi (2011) choose the covariance kernel from a function family such as the squared exponential kernel or Matérn class kernel. However, these covariance functions are defined by measures in a Euclidean space. Here, they depend on both the scalar variables $\boldsymbol{z}_m(t)$ and the functional variables $\boldsymbol{x}_m(\cdot, t)$. To address the problem, we propose a method based on the following new covariance kernel, which allows measures in both Euclidean and functional spaces:

$$\begin{aligned}
&\mathrm{Cov}(\boldsymbol{\tau}_m(\boldsymbol{u}_m(t_1)), \boldsymbol{\tau}_m(\boldsymbol{u}_m(t_2))) = k(\boldsymbol{u}_m(t_1), \boldsymbol{u}_m(t_2)) = k_{\boldsymbol{\theta}}(\boldsymbol{u}_m(t_1), \boldsymbol{u}_m(t_2)) \\
&\ = \theta_{10} \exp\left\{ -\sum_{i=1}^{p} \frac{\theta_{1i}(z_{mi}(t_1) - z_{mi}(t_2))^2}{2} - \sum_{i=1}^{q} \frac{\theta_{1,p+i}||x_{mi}(\cdot, t_1) - x_{mi}(\cdot, t_2)||_\Lambda}{2} \right\} \\
&\quad + \sum_{i=1}^{p} \theta_{2i} z_{mi}(t_1) z_{mi}(t_2) + \sum_{i=1}^{q} \theta_{2,p+i} \int x_{mi}(s, t_1) x_{mi}(s, t_2) ds, \qquad (2.2)
\end{aligned}$$

where $\boldsymbol{\theta} = (\theta_{10}, \theta_{11}, \ldots, \theta_{1Q}, \theta_{21}, \ldots, \theta_{2Q})^\top$ denotes a set of hyper-parameters, with $Q = p + q$, and $||g(\cdot)||_\Lambda$ is a Λ norm of function g. A convenient choice of $||\cdot||_\Lambda$ is the L_2 norm of a function, such as $||g(\cdot)||_\Lambda = \int g(s)^2 ds$. When g belongs to a Hilbert reproducing kernel space with kernel Λ, $||g(\cdot)||_\Lambda$ can be constructed directly; see Appendix A. In (2.2), the first part is a stationary covariance function, which is an extension of the conventional squared exponential function that depends on the generalized distance (Shi and Choi (2011, p. 54)) between two points t_1 and t_2 for a set of mixed scalar and functional variables. We simply use an additive distance here. Other forms of distance may also be used. The remaining part of (2.2) is an extension of a linear covariance function (Shi and Choi (2011, p. 52)), which is nonstationary.

In summary, (1.1) or (2.1) includes a common mean model $c_m(t) = E(y_m(t)| \boldsymbol{u}_m(t)) = \boldsymbol{z}_m^\top(t)\boldsymbol{\nu} + \int_{S_t} \boldsymbol{x}_m^\top(s,t)\boldsymbol{\beta}(s,t)ds$ and random effect τ_m. The latter addresses the problem of heterogeneity, where the covariance structure is defined by our proposed GP priors that allow both scalar and functional covariates. Using similar arguments to those of Shi et al. (2012), this can model the nonlinear relationship between the response variable and the covariates nonparametrically, unlike the mean model $c_m(t)$. At the same time, τ_m are also random effects coping with subject-specific data and characteristics, which can improve predictions for the subject (see Section 4).

2.2. Prediction

From model (1.1), we have the following conditional processes:

$$\begin{gathered} y_m | c_m, \tau_m, \boldsymbol{\theta}, \sigma^2 \sim GP(c_m + \tau_m, \sigma^2 \delta_\epsilon), \\ \tau_m | \boldsymbol{\theta} \sim GP(0, k_{\boldsymbol{\theta}}), \\ y_m | c_m, \boldsymbol{\theta}, \sigma^2 \sim GP(c_m, k_{\boldsymbol{\theta}} + \sigma^2 \delta_\epsilon). \end{gathered}$$

It follows that

$$\begin{gathered} \boldsymbol{y}_m | \boldsymbol{c}_m, \boldsymbol{\tau}_m, \boldsymbol{\theta}, \sigma^2 \sim MVN(\boldsymbol{c}_m + \boldsymbol{\tau}_m, \sigma^2 \boldsymbol{I}), \\ \boldsymbol{\tau}_m | \boldsymbol{\theta} \sim MVN(0, \boldsymbol{K}_m), \\ \boldsymbol{y}_m | \boldsymbol{c}_m, \boldsymbol{\theta}, \sigma^2 \sim MVN(\boldsymbol{c}_m, \boldsymbol{K}_m + \sigma^2 \boldsymbol{I}), \end{gathered} \tag{2.3}$$

where $MVN(b, B)$ stands for multivariate normal distribution with a mean vector b and a covariance matrix B, $\boldsymbol{y}_m = (y_m(t_1), \ldots, y_m(t_n))^\top$ are observations for the mth subject at points $\{t_1, \ldots, t_n\}$, $\boldsymbol{c}_m = (c_m(t_1), \ldots, c_m(t_n))^\top$, $\boldsymbol{\tau}_m = (\tau_m(\boldsymbol{u}_m(t_1)), \ldots, \tau_m(\boldsymbol{u}_m(t_n)))^\top$, $\boldsymbol{K}_m = (k_{\boldsymbol{\theta}}(\boldsymbol{u}_m(t_i), \boldsymbol{u}_m(t_j)))_{n \times n}$, and $\boldsymbol{I}$ is the identity matrix with dimension n.

Denote the data set by $\mathcal{D} = \{(y_m(t_j), \boldsymbol{u}_m(t_j)),\ m = 1, \ldots, M; j = 1, \ldots, n\}$. We now consider a prediction problem in which we assume all parameters are given. The estimation procedure is discussed in the next subsection. Following the derivatives in Appendix A, we have the following posterior distribution of $\boldsymbol{\tau}_m$:

$$\boldsymbol{\tau}_m | \mathcal{D} \sim MVN(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m),$$

where $\boldsymbol{\mu}_m = \boldsymbol{K}_m(\boldsymbol{K}_m + \sigma^2 \boldsymbol{I})^{-1}(\boldsymbol{y}_m - \boldsymbol{c}_m)$, and $\boldsymbol{\Sigma}_m = \boldsymbol{K}_m - \boldsymbol{K}_m(\boldsymbol{K}_m + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{K}_m$.

Now, consider predicting τ_m at a new data point t^*. As shown in Appendix A, we have

$$\tau_m(\boldsymbol{u}_m(t^*)) | \mathcal{D} \sim MVN(\mu_m^*, \sigma_m^*),$$

with $\mu_m^* = \boldsymbol{k}_{mt^*}^\top (\boldsymbol{K}_m + \sigma^2 \boldsymbol{I})^{-1}(\boldsymbol{y}_m - \boldsymbol{c}_m)$, and $\sigma_m^* = k(\boldsymbol{u}_m(t^*), \boldsymbol{u}_m(t^*)) - \boldsymbol{k}_{mt^*}^\top (\boldsymbol{K}_m + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}_{mt^*}$, where $\boldsymbol{k}_{mt} = (k(\boldsymbol{u}_m(t), \boldsymbol{u}_m(t_1)), \ldots, k(\boldsymbol{u}_m(t), \boldsymbol{u}_m(t_n)))^\top$ at time t. Thus, the predictive mean and covariance are $E(y_m(t^*)|\mathcal{D}) = \mu_m^* + c_m(t^*)$ and $Var(y_m(t^*)|\mathcal{D}) = \sigma_m^* + \sigma^2$, respectively. We may use $E(y_m(t^*)|\mathcal{D})$ to estimate

$y_m(t^*)$, denoted by $\hat{y}_m(t^*)$, and use $Var(y_m(t^*)|\mathcal{D})$ to construct the predictive intervals.

Furthermore, we have conditional process $\tau_m|\mathcal{D} \sim GP(\tilde{\mu}_m, \tilde{\sigma}_m)$, where for data points u and v,

$$\begin{aligned}\tilde{\mu}_m(u) &= \boldsymbol{k}_{mu}^\top(\boldsymbol{K}_m + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}_m - \boldsymbol{c}_m),\\ \tilde{\sigma}_m(u,v) &= k(\boldsymbol{u}_m(u), \boldsymbol{u}_m(v)) - \boldsymbol{k}_{mu}^\top(\boldsymbol{K}_m + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}_{mv},\end{aligned}$$

$\boldsymbol{k}_{mu}$ and $\boldsymbol{k}_{mv}$ are $\boldsymbol{k}_{mt}$ at times $t = u$ and v, respectively. Similarly,

$$\begin{aligned}E(y_m(u)|\mathcal{D}) &= \tilde{\mu}_m(u) + c_m(u),\\ Cov(y_m(u), y_m(v)|\mathcal{D}) &= \tilde{\sigma}_m(u,v) + \sigma^2 I(u = v).\end{aligned}$$

We can take $Cov(y_m(\cdot), y_m(\cdot)|\mathcal{D})$ as the estimation of the covariance function for $\hat{y}_m(\cdot)$.

2.3. Parameter estimation

Because $\boldsymbol{\beta}(s,t)$ is a smooth function, it can be approximated by a double expansion in terms of K_s basis functions $\{\phi_k(s), k = 1, \ldots, K_s\}$ and K_t basis functions $\{\psi_l(t), l = 1, \ldots, K_t\}$; that is,

$$\boldsymbol{\beta}(s,t) = \sum_{k=1}^{K_s}\sum_{l=1}^{K_t}\begin{pmatrix} b_{1kl} \\ \vdots \\ b_{qkl}\end{pmatrix}\phi_k(s)\psi_l(t) = \begin{pmatrix}\boldsymbol{\phi}(s)^\top\boldsymbol{B}_1\boldsymbol{\psi}(t) \\ \vdots \\ \boldsymbol{\phi}(s)^\top\boldsymbol{B}_q\boldsymbol{\psi}(t)\end{pmatrix}, \tag{2.4}$$

where $\{b_{ikl}\}$ are coefficients, $\boldsymbol{B}_i = (b_{ikl})_{K_s\times K_t}$, $\boldsymbol{\phi}(s) = (\phi_1(s), \ldots, \phi_{K_s}(s))^\top$, and $\boldsymbol{\psi}(t) = (\psi_1(t), \ldots, \psi_{K_t}(t))^\top$.

Let $\boldsymbol{\phi}_{\boldsymbol{x}mi}(t) = \int_{S_t}\boldsymbol{\phi}(s)x_{mi}(s,t)ds$, which is a vector of length K_s, and

$$\begin{aligned}\boldsymbol{\gamma}_m(t) &= (\boldsymbol{z}_m^\top(t), (\boldsymbol{\psi}(t)\otimes\boldsymbol{\phi}_{\boldsymbol{x}m1}(t))^\top, \ldots, (\boldsymbol{\psi}(t)\otimes\boldsymbol{\phi}_{\boldsymbol{x}mq}(t))^\top)^\top,\\ \boldsymbol{b} &= (\boldsymbol{\nu}^\top, \mathrm{Vec}(\boldsymbol{B}_1)^\top, \ldots, \mathrm{Vec}(\boldsymbol{B}_q)^\top)^\top,\end{aligned}$$

where "⊗" represents the Kronecker product. Hence, $c_m(t) = \boldsymbol{\gamma}_m(t)^\top\boldsymbol{b}$ and $\boldsymbol{c}_m = \boldsymbol{\Gamma}_{mn}^\top\boldsymbol{b}$, where $\boldsymbol{\Gamma}_{mn} = (\boldsymbol{\gamma}_m(t_1), \ldots, \boldsymbol{\gamma}_m(t_n))^\top$. Replacing $c_m(t)$ and $\boldsymbol{c}_m$ with $\boldsymbol{\gamma}_m(t)^\top\boldsymbol{b}$ and $\boldsymbol{\Gamma}_{mn}^\top\boldsymbol{b}$, respectively, in $\hat{y}_m$ and $Var(\hat{y}_m)$, we find that they depend on unknown parameters $\boldsymbol{\theta}$, $\boldsymbol{b}$, and σ^2. We estimate these unknown parameters using a likelihood method.

From (2.3), we have the following marginal density function of $\boldsymbol{y}_m$:

$$\begin{aligned} P(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2) &= \prod_{m=1}^{M} P(\boldsymbol{y}_m|\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2) \\ &= \prod_{m=1}^{M} |2\pi(\boldsymbol{K}_m + \sigma^2 \boldsymbol{I})|^{-1/2} \exp\left\{-\frac{H(\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2)}{2}\right\}, \end{aligned} \tag{2.5}$$

where

$$H(\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2) = (\boldsymbol{y}_m - \boldsymbol{\Gamma}_{mn}^\top \boldsymbol{b})^\top (\boldsymbol{K}_m + \sigma^2 \boldsymbol{I})^{-1} (\boldsymbol{y}_m - \boldsymbol{\Gamma}_{mn}^\top \boldsymbol{b}).$$

For the smoothness of $\boldsymbol{\beta}(\cdot, \cdot)$, it is necessary to add a penalty to the smoothness of the regression function parameter in the log-likelihood function, yielding the objective function,

$$G(\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2) = l(\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2) + \lambda_s \, \texttt{Pen}_s(\boldsymbol{\beta}(s,t)) + \lambda_t \, \texttt{Pen}_t(\boldsymbol{\beta}(s,t)), \tag{2.6}$$

where $l(\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2) = \sum_{m=1}^{M} [\log|\boldsymbol{K}_m + \sigma^2 \boldsymbol{I}| + H(\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2)]$, and λ_s and λ_t are tuning parameters. For penalty functions (Ramsay and Silverman (2005)), we take

$$\texttt{Pen}_s(\boldsymbol{\beta}(s,t)) = \int_a^b \int_a^b \|L_s(\boldsymbol{\beta}(s,t))\|^2 ds dt = \sum_{i=1}^{q} \texttt{trace}[\boldsymbol{B}_i^\top \boldsymbol{L}_{\phi\phi} \boldsymbol{B}_i J_{\psi\psi}],$$

where $\boldsymbol{L}_{\phi\phi} = \int_a^b [L_s \boldsymbol{\phi}(s)][L_s \boldsymbol{\phi}(s)^\top] ds$ and $\boldsymbol{J}_{\psi\psi} = \int_a^b \boldsymbol{\psi}(t) \boldsymbol{\psi}(t)^\top dt$. Similarly,

$$\texttt{Pen}_t(\boldsymbol{\beta}(s,t)) = \int_a^b \int_a^b \|L_t(\boldsymbol{\beta}(s,t))\|^2 ds dt = \sum_{i=1}^{q} \texttt{trace}[\boldsymbol{B}_i^\top \boldsymbol{J}_{\phi\phi} \boldsymbol{B}_i \boldsymbol{L}_{\psi\psi}],$$

where $\boldsymbol{L}_{\psi\psi} = \int_a^b [L_t \boldsymbol{\psi}(t)][L_t \boldsymbol{\psi}(t)^\top] dt$ and $\boldsymbol{J}_{\phi\phi} = \int_a^b \boldsymbol{\phi}(s) \boldsymbol{\phi}(s)^\top ds$.

Solving the derivative of (2.6) with respect to $\boldsymbol{b}$, we obtain the estimation equation,

$$\sum_{m=1}^{M} \boldsymbol{\Gamma}_{mn} (\boldsymbol{K}_m + \sigma^2 \boldsymbol{I})^{-1} \left(\boldsymbol{y}_m - \boldsymbol{\Gamma}_{mn}^\top \boldsymbol{b}\right) = \boldsymbol{\Lambda} \boldsymbol{b},$$

where $\boldsymbol{\Lambda}$ is a $qK_sK_t \times qK_sK_t$ matrix in which the block principal diagonal element is $\lambda_s \boldsymbol{J}_{\psi\psi} \otimes \boldsymbol{L}_{\phi\phi} + \lambda_t \boldsymbol{L}_{\psi\psi} \otimes \boldsymbol{J}_{\phi\phi}$. This yields the following estimator of $\boldsymbol{b}$:

$$\hat{\boldsymbol{b}} = \left[\sum_{m=1}^{M} \boldsymbol{\Gamma}_{mn} (\boldsymbol{K}_m + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{\Gamma}_{mn}^\top + \boldsymbol{\Lambda}\right]^{-1} \sum_{m=1}^{M} \boldsymbol{\Gamma}_{mn} (\boldsymbol{K}_m + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}_m. \tag{2.7}$$

If $\tau_m = 0$ (i.e., the function-on-function linear model without random effects), the estimate of $\boldsymbol{b}$ is also given by (2.7), but with $\boldsymbol{K}_m = 0$.

We can estimate $\boldsymbol{\theta}$ and σ similarly. The estimation procedure is as follows.

Algorithm 1

```
For initial values $\boldsymbol{\theta} = \boldsymbol{\theta}^*$,
(I) Given $\boldsymbol{\theta}$, we update estimates of $\boldsymbol{b}, \sigma^2$ by
    $$\hat{\boldsymbol{b}}, \hat{\sigma}^2 = \underset{\boldsymbol{b},\sigma}{\operatorname{argmin}}\, G(\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2). \quad (2.8)$$
(II) Given $\boldsymbol{b} = \hat{\boldsymbol{b}}, \sigma^2 = \hat{\sigma}^2$, we estimate $\boldsymbol{\theta}$ by
    $$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{m=1}^{M} [\log|\boldsymbol{K}_m + \sigma^2 \boldsymbol{I}| + H(\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2)]. \quad (2.9)$$
(III) Repeat (I) and (II) until some convergence criteria is met.
```

We employ the log-likelihood, $-l(\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2)/2$, to construct the convergence criterion. When the absolute relative difference $l(\boldsymbol{\theta}, \boldsymbol{b}, \sigma^2)$ between two successive iterations is less than a prespecified value, the iteration procedure stops, and the parameter estimates and random-effect prediction are calculated.

3. Asymptotic Properties

The common mean structure is estimated using data collected from all M subjects, and has been proved consistent in many functional linear models under suitable regularity conditions; see Yao, Müller and Wang (2005b), Yuan and Cai (2010), Sun et al. (2018), and others. Information consistency reflects whether a prediction $\hat{y}_m(t)$ converges to its true curve $y_m(t)$ when we have enough data collected from the mth subject. For the GPR model and extended T-process model, the properties are obtained in Seeger, Kakade and Foster (2008), Wang and Shi (2014), and Wang, Shi and Lee (2017). Here, we show this property holds for model (1.1) as well.

Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, where $\mathcal{X}_1$ and $\mathcal{X}_2$ are the spaces to which covariates $\boldsymbol{z}_m(t)$ and $\boldsymbol{x}_m(\cdot, t)$. From the true model (2.1) and the assumed model (1.1), let $p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m}, \boldsymbol{u}_m)$ be the true density function of a $\boldsymbol{y}_m$, and

$$p_{\sigma,\theta}(\boldsymbol{y}_m|\boldsymbol{u}_m) = \int_{\mathcal{F}} p_\sigma(\boldsymbol{y}_m|\tau, \boldsymbol{u}_m) dp_\theta(\tau),$$

where τ_{0m} is the true underlying function of τ_m, σ_0 is the true value of σ, and

$p_\theta(\tau)$ is a measure of random process τ on space $\mathcal{F} = \{\tau(\cdot,\cdot) : \mathcal{X} \to R\}$. In Appendix B, we show that

$$p_{\sigma,\theta}(\boldsymbol{y}_m|\boldsymbol{u}_m) = \prod_{l=1}^{n} p_{\sigma,\theta}(y_{ml}|\boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)}), \tag{3.1}$$

where

$$\begin{aligned} p_{\sigma,\theta}(y_{ml}|\boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)}) &= \int_{\mathcal{F}} p_\sigma(y_{ml}|\tau, \boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)}) dp_\theta(\tau|\boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)}), \\ p_\theta(\tau|\boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)}) &= \frac{p_\sigma(\boldsymbol{y}_{m(l-1)}|\tau, \boldsymbol{u}_{m(l-1)}) p_\theta(\tau)}{\int_{\mathcal{F}} p_\sigma(\boldsymbol{y}_{m(l-1)}|\tau', \boldsymbol{u}_{m(l-1)}) dp_\theta(\tau')}, \end{aligned}$$

$\boldsymbol{y}_{ml} = (y_{m1}, \ldots, y_{ml})^\top$, and $\boldsymbol{u}_{ml} = \{\boldsymbol{z}_{m1}, \ldots, \boldsymbol{z}_{ml}, \boldsymbol{x}_{m1}, \ldots, \boldsymbol{x}_{ml}\}$, for $l = 1, \ldots, n$. Under the true model (2.1), we also have

$$p_\sigma(\boldsymbol{y}_m|\tau_{0m}, \boldsymbol{u}_m) = \prod_{l=1}^{n} p_\sigma(y_{ml}|\tau_{0m}, \boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)}). \tag{3.2}$$

Here, $p_\sigma(y_{ml}|\tau_{0m}, \boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)})$ and $p_{\sigma,\theta}(y_{ml}|\boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)})$ can be treated as Bayesian predictive models (Seeger, Kakade and Foster (2008)).

Let $p_{\sigma_0,\hat{\theta}}(\boldsymbol{y}_m|\boldsymbol{x}_m)$ be the estimated density function under the assumed model (1.1), where $\hat{\boldsymbol{\theta}}$ is the estimator of parameter $\boldsymbol{\theta}$. Denote $D[p_1, p_2] = \int (\log p_1 - \log p_2) dp_1$ as the Kullback−Leibler divergence between two densities p_1 and p_2. It follows from (3.1) and (3.2) that

$$\begin{aligned} & D[p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m}, \boldsymbol{u}_m), p_{\sigma_0,\hat{\theta}}(\boldsymbol{y}_m|\boldsymbol{u}_m)] \\ &= \int \sum_{l=1}^{n} Q(y_{ml}|\boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)}) p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m}, \boldsymbol{u}_m) d\boldsymbol{y}_m, \end{aligned}$$

where $Q(y_{ml}|\boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)}) = \log\{p_{\sigma_0}(y_{ml}|\tau_{0m}, \boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)})/p_{\sigma_0,\hat{\theta}}(y_{ml}|\boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)})\}$ is a loss function, and $\sum_{l=1}^{n} Q(y_{ml}|\boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)})$ is called a cumulative loss. Assuming that the mean in (1.1) and $\boldsymbol{\theta}$ are consistent, we prove in Appendix C that the average cumulative loss tends to zero asymptotically. The result is presented in the following theorem.

Theorem 1. *Under the appropriate conditions specified in Appendix C, we have*

$$\frac{1}{n} E_{\boldsymbol{u}_m}(D[p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m}, \boldsymbol{u}_m), p_{\sigma_0,\hat{\theta}}(\boldsymbol{y}_m|\boldsymbol{u}_m)]) \longrightarrow 0, \textit{as} \quad n \to \infty,$$

where the expectation is taken over the distribution of $\boldsymbol{u}_m$.

Theorem 1 shows that the Kullback−Leibler divergence between two density functions for $(\boldsymbol{y}_m|\boldsymbol{u}_m)$ from the true and the assumed models, respectively, tends to zero, asymptotically.

4. Numerical Studies

4.1. Simulation studies

Simulation studies were conducted to evaluate the performance of the proposed model (M2) by comparing it with that of the common functional linear model, ignoring random effects (M1). Data are generated from the following model:

$$y_m(t) = z_m(t)\nu + \int_0^1 x_m(s,t)\beta(s,t)ds + \tau_m(z_m(t), x_m(\cdot,t)) + \epsilon_m(t), \;\; m = 1,\ldots,M, \tag{4.1}$$

where $z_m(\cdot)$ is a GP with mean function $h_1(t) = t$, for $t \in (0,1)$, and kernel function

$$k_1(z_m(t_1)), z_m(t_2))) = g(t_1,t_2) = 0.1\exp\{-5(t_1-t_2)^2\} + 0.1t_1t_2,$$

where $x_m(\cdot,\cdot)$ is generated from a GP with mean function $h_2(s,t) = t + \cos(s)$, for $s,t \in (0,1)$, and kernel function $k_2(x_m(s_1,t)), x_m(s_2,t))) = g(s_1,s_2)$. We consider six different combinations of τ_m and $\beta(s,t)$:

S1: τ_m is a GP with mean zero and the kernel function (2.2), and $\beta(s,t) = (t^2+\cos(s))/10$, for $s,t \in (0,1)$;

S2: τ_m is a GP with mean zero and the kernel function (2.2), and $\beta(s,t) = \exp\{-(t^2+s^2)\}/10$, for $s,t \in (0,1)$;

S3: τ_m is a GP with mean zero and the kernel function (2.2), and $\beta(s,t) = (-t^2+\cos(s))/10$, for $s,t \in (0,1)$;

S4: $\tau_m = 0$ and $\beta(s,t) = (t^2+\cos(s))/10$, for $s,t \in (0,1)$;

S5: $\tau_m = 0$ and $\beta(s,t) = \exp\{-(t^2+s^2)\}/10$, for $s,t \in (0,1)$;

S6: $\tau_m = 0$ and $\beta(s,t) = (-t^2+\cos(s))/10$, for $s,t \in (0,1)$;

where $\nu = 1.0$, $\theta_{10} = \theta_{12} = \theta_{21} = \theta_{22} = 0.1, \theta_{11} = 10$, and $\sigma^2 = 0.5$. Under S1, S2, and S3, model M2 (model (1.1)) is the true model, whereas M1 holds for S4, S5, and S6. Sample size $M = 10, 20$, and 30. Both t and s take 20 points, equally spaced in (0, 1), where the data at 10 points of t are chosen as training data, and the remainder are used as test data. All simulations are repeated 500 times.

The prediction errors are given by $f_0(t) - \hat{f}(t)$, where $f_0(t) = z_m(t)\nu_0 + \int_0^1 x_m(s,t)\beta_0(s,t)ds + \tau_{0m}(z_m(t), x_m(\cdot,t))$ and $\hat{f}(t) = z_m(t)\hat{\nu} + \int_0^1 x_m(s,t)\hat{\beta}(s,t)ds + \hat{\tau}_m(z_m(t), x_m(\cdot,t))$. From model (4.1), $f_0(t)$ is the true regression function and $\hat{f}(t)$ is an estimator of $f_0(t)$. Figure 1 plots the prediction errors using M1 and M2 under Cases S3 and S6. We see that M1 and M2 for $\tau_m = 0$ (S6) have comparable results. However, when $\tau_m \neq 0$ (S3), M2 has much smaller prediction errors than those of M1.

The actual performance can be measured by the following summary statistics: prediction error (PE), $PE = \sum_{i=1}^{M}\sum_{k=1}^{n}(y(t_i) - \hat{f}(t_i))^2/(nM)$ and average bias (AB), $AB = \sum_{i=1}^{M}\sum_{k=1}^{n}(\hat{f}(t_i) - f_0(t_i))^2/(nM)$. Tables 1 and 2 present the values of PE and AB for predictions based on the training data and the test data, respectively. It shows that when $\tau_m = 0$ (S4, S5, and S6), where M1 is the true model, PE and AB from M1 and M2 are comparable. However, under S1, S2, and S3, the predictions based on M2 have much smaller PE and AB than those from M1. As the sample size increases, PE and AB from M2, and their standard errors (SD), become smaller. In conclusion, the proposed M2 exhibits comparable performance with M1 when there is no heterogeneity among subjects, but M2 performs much better than M1 when heterogeneity does exist.

4.2. Real-data examples

4.2.1. Canadian weather data

Canadian weather data consist of temperature and precipitation at 35 Canadian weather observation stations, and are available in the R-package fda (Ramsay, Hooker and Graves (2010)). These stations are divided into four regions: Arctic, Atlantic, Continental, and Pacific. For each region, we assume that there is a common cumulative, lagged temperature effect to precipitation, and investigate the functional relationship between the two. Owing to the spatial nature of the weather data, we consider heterogeneity among the stations. Hence, we investigate the data using our flexible functional regression model, thus allowing for both cumulative, lagged temperature effects and the spatial correlation structure among weather stations; that is,

$$y_{ij}(t) = a + \int_0^t x_{ij}(s)\beta_i(s,t)ds + \tau_{ij}(x_{ij}(\cdot,t)) + \epsilon_{ij}(t), \tag{4.2}$$

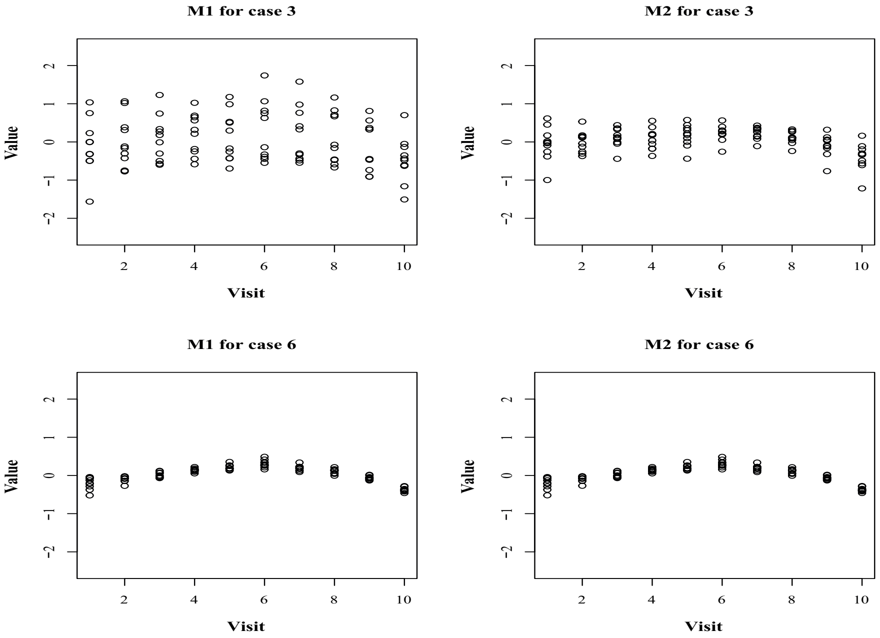


Figure 1. Prediction errors for two methods without random effects (M1) and with random effects (M2) under Cases 3 and 6.

Table 1. PE and AB using M1 (without random effects) and M2 (with random effects) for the training data, where SDs are presented in parentheses.

		$M = 10$		$M = 20$		$M = 30$	
Model	method	PE	AB	PE	AB	PE	AB
S1	M1	0.565(0.141)	0.344(0.121)	0.603(0.102)	0.367(0.091)	0.616(0.086)	0.375(0.076)
	M2	0.273(0.058)	0.115(0.028)	0.274(0.043)	0.107(0.019)	0.274(0.041)	0.104(0.016)
S2	M1	0.529(0.137)	0.308(0.119)	0.567(0.1)	0.33 (0.09)	0.579(0.085)	0.337(0.075)
	M2	0.244(0.051)	0.085(0.025)	0.243(0.037)	0.075(0.015)	0.244(0.033)	0.073(0.012)
S3	M1	0.567(0.139)	0.344(0.12)	0.605(0.101)	0.367(0.091)	0.604(0.082)	0.363(0.074)
	M2	0.272(0.06)	0.114(0.029)	0.273(0.043)	0.105(0.019)	0.276(0.041)	0.104(0.016)
S4	M1	0.281(0.04)	0.054(0.008)	0.286(0.028)	0.048(0.005)	0.287(0.024)	0.046(0.003)
	M2	0.28 (0.04)	0.053(0.008)	0.285(0.027)	0.047(0.005)	0.285(0.024)	0.044(0.003)
S4	M1	0.243(0.036)	0.016(0.008)	0.249(0.024)	0.011(0.004)	0.249(0.021)	0.008(0.002)
	M2	0.243(0.036)	0.016(0.008)	0.249(0.024)	0.01 (0.004)	0.249(0.021)	0.008(0.002)
S6	M1	0.279(0.042)	0.053(0.008)	0.286(0.028)	0.048(0.005)	0.286(0.023)	0.045(0.003)
	M2	0.278(0.042)	0.052(0.008)	0.285(0.028)	0.046(0.005)	0.284(0.023)	0.043(0.003)

Table 2. PE and AB using M1 (without random effects) and M2 (with random effects) for the test data, where SDs are presented in parentheses.

		$M = 10$		$M = 20$		$M = 30$	
Model	method	PE	AB	PE	AB	PE	AB
S1	M1	0.677(0.161)	0.424(0.137)	0.688(0.116)	0.44 (0.101)	0.693(0.095)	0.446(0.083)
	M2	0.43 (0.065)	0.178(0.043)	0.415(0.046)	0.164(0.028)	0.411(0.038)	0.161(0.023)
S2	M1	0.597(0.156)	0.345(0.133)	0.611(0.111)	0.362(0.098)	0.614(0.094)	0.367(0.081)
	M2	0.353(0.054)	0.103(0.032)	0.341(0.037)	0.091(0.018)	0.337(0.029)	0.087(0.014)
S3	M1	0.677(0.162)	0.426(0.138)	0.691(0.115)	0.442(0.101)	0.683(0.092)	0.433(0.082)
	M2	0.428(0.066)	0.177(0.043)	0.417(0.046)	0.165(0.028)	0.41 (0.039)	0.16 (0.023)
S4	M1	0.355(0.055)	0.104(0.028)	0.346(0.037)	0.095(0.017)	0.344(0.03)	0.092(0.013)
	M2	0.354(0.054)	0.103(0.028)	0.345(0.037)	0.094(0.017)	0.342(0.03)	0.09 (0.013)
S5	M1	0.275(0.042)	0.024(0.013)	0.267(0.028)	0.017(0.008)	0.265(0.023)	0.015(0.006)
	M2	0.275(0.042)	0.024(0.013)	0.267(0.028)	0.017(0.008)	0.264(0.023)	0.015(0.006)
S6	M1	0.355(0.057)	0.104(0.027)	0.346(0.038)	0.097(0.017)	0.342(0.03)	0.094(0.014)
	M2	0.354(0.057)	0.103(0.027)	0.344(0.038)	0.095(0.017)	0.34 (0.03)	0.091(0.014)

where $y_{ij}(t)$ and $x_{ij}(t)$ denote the precipitation and the temperature, respectively, for the jth station at region i and time t. This model is a special case of model (1.1) with covariates $z_{ij}(t) = 1$ and $x_{ij}(s,t) = x_{ij}(s)$. Before applying our method to this data, the precipitation values need to be standardized.

Figure 2 presents the predictions using M1 and M2 in the region Arctic, that is, assuming fixed and random effects, respectively (M2 is the proposed model with random effects, whereas M1 has only fixed effects). Figure 3 shows the estimations of the fixed and random effects for the other three regions: Atlantic, Continental and Pacific. Figure 2 shows that M2 provides quite different results to those from M1. M2 has an average squared residual (ASR) of 0.215, whereas M1 has an ASR of 0.499, suggesting that M2 fits the data much better than M1 does. We see that precipitation for each station in the four regions has different random effects. For example, in the Arctic, three weather stations have very different effects, which is reasonable, because those stations are far apart. Cross-validation is also used to show the performance of the prediction. The mean squares of the prediction errors for 10-fold cross validation are 0.722 and 0.314 for the methods M1 and M2, respectively. Thus, the proposed method M2 is more accurate than M1.

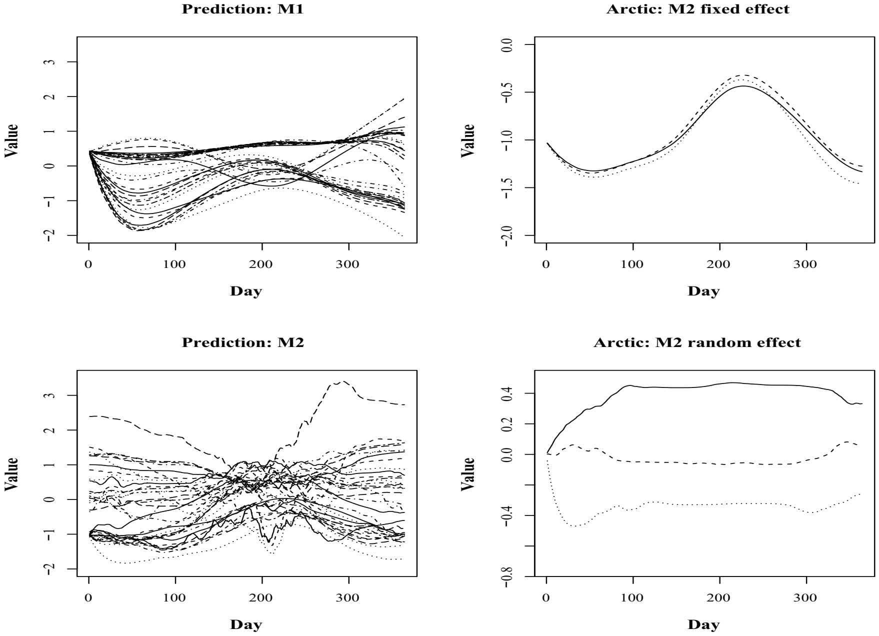


Figure 2. Predictions for two models without random effects (M1) and with random effects (M2), and fixed and random effects for Arctic.

4.2.2. Movement data

The data were collected from 70 stroke survivors, consisting of 34 acute patients with a stroke less than one month previously, and 36 chronic patients with a stroke more than six months previously. The response variable, the dependency level of the patients in their daily life or the impairment level of their upper-limb function, is measured using Chedoke Arm and Hand Activity Inventory, or CAHAI (`http://www.cahai.ca/`). Each patient has up to eight scheduled assessments over three months, where the first assessment provides a baseline level. Details can be found in Shi et al. (2013) and Serradilla et al. (2014).

We focus on acute patients, where there are 173 observations from 34 patients, with 72 functional variables from 10 movements, and 68 kinematic scalar variables from 17 movements. Based on the discussion in Cheng, Shi and Eyre (2017), we employ three bivariate functional variables: forward circle movement of paretic limb from x-axis (x_{m1} = LA05.lx), sawing movement of paretic limb from y-axis (x_{m2} = LA09.ly), and orientation movement of nonparetic limb from x-axis (x_{m3} = LA28.rqx). In addition, we consider three kinematic scalar variables: the speeds of the paretic limb for the forward circle at movements LA05,

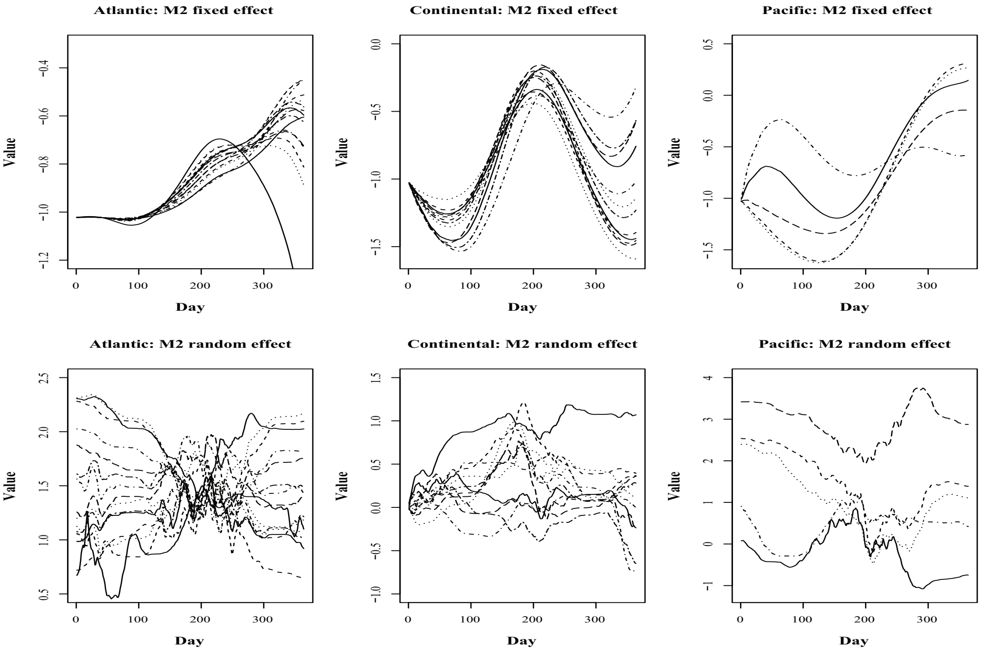


Figure 3. Fixed and random effects for three regions: Atlantic, Continental and Pacific.

LA10, and LA34 (z_{m1} = sp_P_LA05, z_{m2} = sp_P_LA10, and z_{m3} = sp_P_LA34). The impairment level is standardized before analysis.

We consider the functional relationship between the impairment level and the functional movement variables. At each assessment time, the movement variables (i.e., forward circle, sawing, and orientation), are curves; thus, investigating the cumulative effects on movements of an impairment level is essential. Owing to personal healthy status, we also study the heterogeneity effect for each patient. Hence, we rewrite model (1.1) as

$$y_m(t)=\boldsymbol{z}_m^{\top}(t)\boldsymbol{\nu}+\int_0^{T_0}\boldsymbol{x}_m^{\top}(s,t)\boldsymbol{\beta}(s,t)ds+\tau_m(\boldsymbol{z}_m(t),\boldsymbol{x}_m(\cdot,t))+\epsilon_m(t),\ m=1,\ldots,34,$$

where $\boldsymbol{z}_m(t)=(z_{m1}(t),z_{m2}(t),z_{m3}(t))^{\top}$ and $\boldsymbol{x}_m(\cdot,t)=(x_{m1}(\cdot,t),x_{m2}(\cdot,t),x_{m3}(\cdot,t))^{\top}$, and T_0 is a prespecified end time for the measurement of each movement.

Figure 4 presents the predictions from two estimated models M1 and M2, that is, with fixed and random effects, respectively. The findings are similar to those for the Canadian weather data. The residuals from M2 (ASR of 0.076) are much smaller than those from M1 (ASR of 0.479), showing that M2 is more

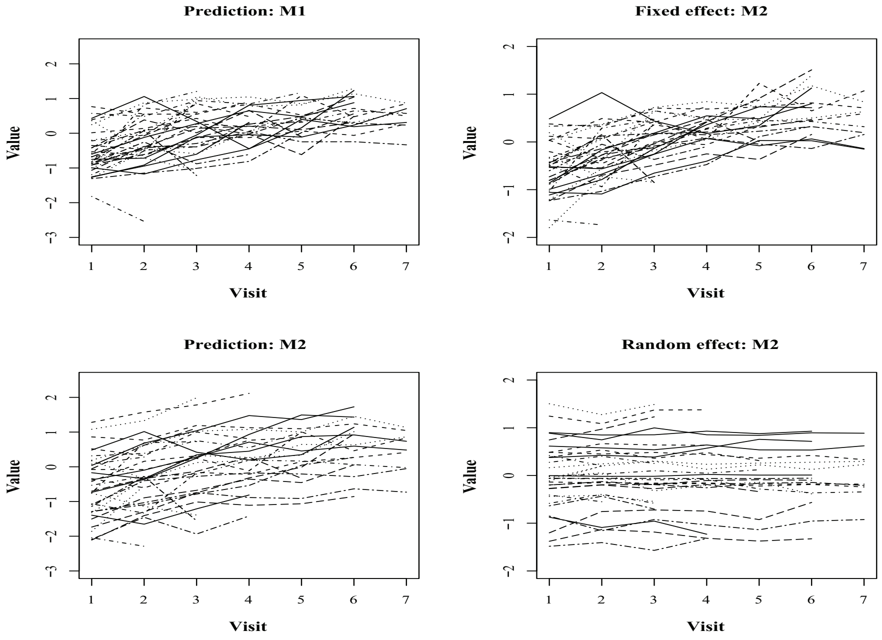


Figure 4. Predictions for two models without random effects (M1) and with random effects (M2).

suitable for fitting the movement data than M1. Again, the random effects for each stroke survivor are different. Figure 5 presents the profile plots of $\hat{\beta}_i(s,t)$, for visiting time $t = 1$, 3, 5 and 7. 3D plots of $\hat{\beta}_i(s,t)$ are presented in Appendix D. The shapes of $\hat{\beta}_i(s,t)$ against visiting time t for the forward circle and sawing movements of the paretic limb are quite different to those of the orientation movement of the nonparetic limb. This makes sense because the first two movements are conducted by the paretic limb, whereas the third by the health limb, showing the longitudinal effects of the paretic side and the fixed effects of the nonparetic side. This may also indicate the improvement in the function of the upper limbs after using the home-based rehabilitation system for a few months.

The scalar variables, corresponding to the speeds of the paretic limb, have parameter estimation $\hat{\boldsymbol{\nu}}^\top = (0.347, 0.084, 0.693)^\top$, which suggests that the CA-HAI score becomes higher when the patient can move the paretic limb faster. The mean square of the prediction errors using cross-validation are 0.269 and 0.591 (24.314 and 53.323 under original scale of response) for the methods M1 and M2, respectively. Again, M2 outperforms M1.

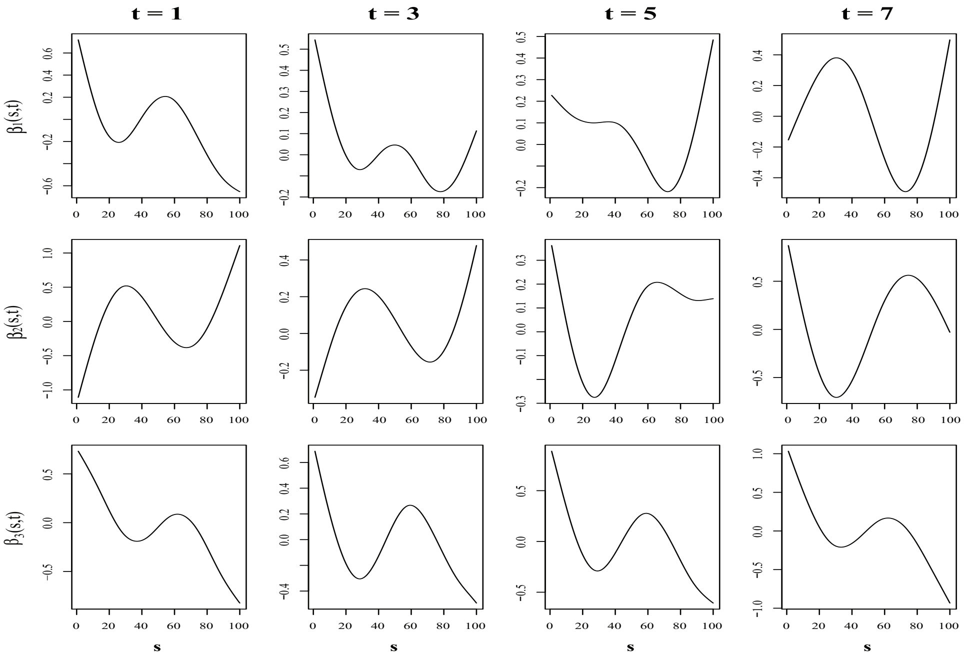


Figure 5. Estimations of functions $\beta_i(s,t)$, for $i = 1, 2, 3$, and visiting time $t = 1, 3, 5, 7$.

5. Conclusion

This paper proposes a nonparametric random effects functional regression model using GP priors and develops a flexible and accurate procedure to calculate estimation and prediction. This model builds a flexible framework, coping with different types of mean models, for example the function-on-function linear model, historical functional regression model, and concurrent functional linear model. In addition, the model introduces nonparametric random effects using GP priors, allowing the use of subject-specific data and characteristics, thus improving model fitting and prediction in many cases. Information consistency of the prediction has been proved. Numerical studies show that the proposed model outperforms models that do not assume random effects. We focused our discussion on the error term with Gaussian distribution in this paper, but the estimation procedure can be extended to generalized models with functional data (Wang and Shi (2014)). The GP priors may also be replaced by other process priors, for example, robust heavy-tailed processes (Wang, Shi and Lee (2017); Cao, Shi and Lee (2018)). However, such an extension may be not straightforward, because there are no closed forms for the parameter estimation and prediction of a random effect when a heavy-tailed process prior, (e.g., t-process), is used. These topics are left to future research.

Appendices

A. Kernel Function and Prediction Distributions

Kernel function

If g belongs to a Hilbert reproduce kernel space with kernel Λ, then by Mercer's theorem, Λ can be decomposed as $\Lambda(s,t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(s)\phi_i(t)$, where λ_i and ϕ_i are eigenvalue and eigenfunction of Λ. From this decomposition, we have $g(t) = \sum_{i=1}^{\infty} \xi_i \phi_i(t)$ with $\xi_i = \int g(s)\phi_i(s)ds$. Hence,

$$||g(\cdot)||_\Lambda =< \sum_{i=1}^{\infty} \xi_i\phi_i, \sum_{i=1}^{\infty} \xi_i\phi_i >_\Lambda = \sum_{i=1}^{\infty} \frac{\xi_i^2}{\lambda_i} < \phi_i, \phi_i > = \sum_{i=1}^{\infty} \frac{\xi_i^2}{\lambda_i},$$

where $< \cdot, \cdot >_\Lambda$ and $< \cdot, \cdot >$ stand for inner products under Λ and L_2, respectively. If we observed g at data points $v_i, i = 1, \ldots, I$, then by Representer Theorem, $g(t)$ can be approximated with $g(t) = \sum_{i=1}^{I} a_i \Lambda(v_i, t)$ which leads to

$$\begin{aligned}
||g(\cdot)||_\Lambda &=< \sum_{i=1}^{I} a_i\Lambda(v_i,\cdot), \sum_{i=1}^{I} a_i\Lambda(v_i,\cdot) >_\Lambda = \sum_{i=1}^{I}\sum_{j=1}^{I} a_i a_j < \Lambda(v_i,\cdot), \Lambda(v_j,\cdot) >_\Lambda \\
&= \sum_{i=1}^{I}\sum_{j=1}^{I} a_i a_j < \sum_{k=1}^{\infty} \lambda_k\phi_k(v_i)\phi_k(\cdot), \sum_{k=1}^{\infty} \lambda_k\phi_k(v_j)\phi_k(\cdot) >_\Lambda \\
&= \sum_{i=1}^{I}\sum_{j=1}^{I} a_i a_j \sum_{k=1}^{\infty} \lambda_k^2\phi_k(v_i)\phi_k(v_j) < \phi_k(\cdot), \phi_k(\cdot) >_\Lambda \\
&= \sum_{i=1}^{I}\sum_{j=1}^{I} a_i a_j \sum_{k=1}^{\infty} \lambda_k\phi_k(v_i)\phi_k(v_j) = \sum_{i=1}^{I}\sum_{j=1}^{I} a_i a_j \Lambda(v_i, v_j).
\end{aligned}$$

Prediction distributions

From model (2.3), we have

$$\begin{pmatrix} \boldsymbol{y}_m \\ \boldsymbol{\tau}_m \end{pmatrix} \Bigg| \boldsymbol{u}_m \sim MVN\left((\boldsymbol{c}_m^\top, 0)^\top, \begin{pmatrix} \boldsymbol{K}_m + \sigma^2\boldsymbol{I} & \boldsymbol{K}_m \\ \boldsymbol{K}_m & \boldsymbol{K}_m \end{pmatrix} \right), \tag{A.1}$$

with the same notations defined after (2.3). From derivative of conditional normal distribution, the join distribution (A.1) indicates that

$$\boldsymbol{\tau}_m|\mathcal{D} \sim MVN(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m),$$

where $\boldsymbol{\mu}_m = \boldsymbol{K}_m(\boldsymbol{K}_m + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}_m - \boldsymbol{c}_m), \boldsymbol{\Sigma}_m = \boldsymbol{K}_m - \boldsymbol{K}_m(\boldsymbol{K}_m + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{K}_m$.

Again, from model (1.1), we obtain

$$\begin{pmatrix} \boldsymbol{y}_m \\ \tau_m(\boldsymbol{u}_m(t^*)) \end{pmatrix} \Bigg| \boldsymbol{u}_m \sim MVN\left((\boldsymbol{c}_m^\top, 0)^\top, \begin{pmatrix} \boldsymbol{K}_m + \sigma^2\boldsymbol{I} & \boldsymbol{k}_{mt^*}^\top \\ \boldsymbol{k}_{mt^*} & k(\boldsymbol{u}_m(t^*), \boldsymbol{u}_m(t^*)) \end{pmatrix} \right),$$

where $\boldsymbol{k}_{mt^*} = (k(\boldsymbol{u}_m(t^*), \boldsymbol{u}_m(t_1)), \ldots, k(\boldsymbol{u}_m(t^*), \boldsymbol{u}_m(t_n)))^\top$. It follows that $\tau_m(\boldsymbol{u}_m(t^*))|\mathcal{D} \sim MVn(\mu_m^*, \sigma_m^*)$, with

$$\mu_m^* = \boldsymbol{k}_{mt^*}^\top(\boldsymbol{K}_m + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}_m - \boldsymbol{c}_m),$$
$$\sigma_m^* = k(\boldsymbol{u}_m(t^*), \boldsymbol{u}_m(t^*)) - \boldsymbol{k}_{mt^*}^\top(\boldsymbol{K}_m + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}_{mt^*}.$$

B. Proof of (3.1)

From Bayes' Theorem, we have

$$\begin{aligned}
&\prod_{l=1}^{n} p_{\sigma,\theta}(y_{ml}|\boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)}) \\
&= p_{\sigma,\theta}(y_{m1}|\boldsymbol{u}_{m1}) \prod_{l=2}^{n} \int_{\mathcal{F}} p_\sigma(y_{ml}|\tau, \boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)}) dp_\theta(\tau|\boldsymbol{u}_{ml}, \boldsymbol{y}_{m(l-1)}) \\
&= p_{\sigma,\theta}(y_{m1}|\boldsymbol{u}_{m1}) \prod_{l=2}^{n} \int_{\mathcal{F}} \frac{p_\sigma(\boldsymbol{y}_{ml}|\tau, \boldsymbol{u}_{ml}) dp_\theta(\tau)}{\int_{\mathcal{F}} p_\sigma(\boldsymbol{y}_{m(l-1)}|\tau', \boldsymbol{u}_{m(l-1)}) dp_\theta(\tau')} \\
&= \int_{\mathcal{F}} p_\sigma(\boldsymbol{y}_m|\tau, \boldsymbol{u}_m) dp_\theta(\tau) = p_{\sigma,\theta}(\boldsymbol{y}_m|\boldsymbol{u}_m),
\end{aligned}$$

which shows that the equation (3.1) holds.

C. Information Consistency

Lemma 1. *Suppose $\boldsymbol{y}_m$ are independently sampled from (2.1) and $\tau_{0m} \in \mathcal{F}$ follow a Gaussian process $GP(0, k)$ with bounded function $k(\cdot, \cdot; \boldsymbol{\theta})$ for any covariances values in $\mathcal{X}$. Suppose $k(\cdot, \cdot; \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$, then, we have*

$$-\log\{p_{\sigma_0,\hat{\boldsymbol{\theta}}}(\boldsymbol{y}_m|\boldsymbol{u}_m)\} + \log\{p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m}, \boldsymbol{u}_m)\}$$

$$\leq c + \frac{1}{2}\log|\boldsymbol{I} + \sigma^{-2}\boldsymbol{K}_m| + \frac{1}{2}\|\tau_{0m}\|_k^2 + \delta \tag{C.1}$$

where $\|\tau_{0m}\|_k$ is the reproducing kernel Hilbert space(RKHS) norm of τ_{0m} associated with $k(\cdot,\cdot;\boldsymbol{\theta})$, $\boldsymbol{K}_m$ is covariance matrix of τ_{0m} over $\boldsymbol{u}_m$, $\boldsymbol{I}$ is the $n \times n$ identity matrix and c, δ are some positive constants.

Proof: Denoted by $\mathcal{H}$ the reproducing kernel Hilbert space(RKHS) associated with $k(\cdot,\cdot;\boldsymbol{\theta})$, and let $\mathcal{H}_n = \{f(\cdot) : f(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i k(\boldsymbol{x}, \boldsymbol{u}_i;\boldsymbol{\theta}), \text{for any } \alpha_i \in R\}$ be the span of $\{k(\cdot,\boldsymbol{u}_i;\boldsymbol{\theta}\}$. First, we assume that $\tau_{0m} \in \mathcal{H}_n$, then we have

$$\tau_{0m}(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot,\boldsymbol{u}_i;\boldsymbol{\theta}) \triangleq K(\cdot)\boldsymbol{\alpha}, \tag{C.2}$$

where $K(\cdot) = (k(\cdot,\boldsymbol{u}_1;\boldsymbol{\theta}),\ldots,k(\cdot,\boldsymbol{u}_n;\boldsymbol{\theta}))$ and $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_n)^\top$. With the property of RKHS, $(\tau_{0m}(\boldsymbol{u}_1),\ldots,\tau_{0m}(\boldsymbol{u}_n))^\top = \boldsymbol{K}_m\boldsymbol{\alpha}$ and $\|\tau_{0m}\|_k^2 = \boldsymbol{\alpha}^\top\boldsymbol{K}_m\boldsymbol{\alpha}$.

Let $D[p_1,p_2] = \int(\log p_1 - \log p_2)dp_1$ be the Kullback-Leibler distance between two densities p_1 and p_2. By the Fenchel-Legendre duality relationship, we have

$$-\log p_{\sigma_0,\theta}(\boldsymbol{y}_m|\boldsymbol{u}_m) \leq -E_Q\{\log p(\boldsymbol{y}_m|\tau_m,\boldsymbol{u}_m)\} + D[Q,P], \tag{C.3}$$

where P is the measure induced by $GP(0,k(\cdot,\cdot;\hat{\boldsymbol{\theta}}))$, Q is the posterior distribution of $\boldsymbol{\tau}_m$ with a prior distribution $GP(0,k(\cdot,\cdot;\hat{\boldsymbol{\theta}}))$ and normal likelihood $\prod_{i=1}^n N(\hat{y_i};\tau_m(u_{mi}),\sigma^2)$, $(\hat{y_1},\ldots,\hat{y_n})^\top = (\boldsymbol{K}_m + \sigma^2\boldsymbol{I})\boldsymbol{\alpha}$. Then the posterior distribution Q of $\boldsymbol{\tau}_m$ is $N(\boldsymbol{K}_m\boldsymbol{\alpha}, \boldsymbol{K}_m\boldsymbol{B}^{-1})$, where $\boldsymbol{B} = \boldsymbol{I} + \sigma^{-2}\boldsymbol{K}_m$. Then it easily shows that

$$\begin{aligned} D[Q,P] = \frac{1}{2}\Big\{ &-\log|\hat{\boldsymbol{K}}_m^{-1}\boldsymbol{K}_m| + \log|\boldsymbol{B}| + \text{tr}(\hat{\boldsymbol{K}}_m^{-1}\boldsymbol{K}_m\boldsymbol{B}^{-1}) \\ &+ \|\tau_{0m}\|_k^2 + \boldsymbol{\alpha}^\top\boldsymbol{K}_m(\hat{\boldsymbol{K}}_m^{-1}\boldsymbol{K}_m - \boldsymbol{I})\boldsymbol{\alpha} - n\Big\}, \end{aligned} \tag{C.4}$$

where $\hat{\boldsymbol{K}}_m$ is defined in the same way as $\boldsymbol{K}_m$ but with $\boldsymbol{\theta}$ being replaced by its estimator $\hat{\boldsymbol{\theta}}$. On the other hand,

$$E_Q\{\log p(\boldsymbol{y}_m|\tau_m,\boldsymbol{u}_m)\} \geq \log p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m},\boldsymbol{u}_m) - \frac{1}{2\sigma^2}\text{tr}(\boldsymbol{K}_m\boldsymbol{B}^{-1}). \tag{C.5}$$

Hence, it follow from (C.3), (C.4) and (C.5) that

$$\begin{aligned} &-\log p_{\sigma_0,\theta}(\boldsymbol{y}_m|\boldsymbol{u}_m) + \log p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m},\boldsymbol{u}_m) \\ &\leq \frac{1}{2}\Big\{ -\text{log}|\hat{\boldsymbol{K}}_m^{-1}\boldsymbol{K}_m| + \text{log}|\boldsymbol{B}| + \text{tr}\left(\left(\hat{\boldsymbol{K}}_m^{-1}\boldsymbol{K}_m + \frac{1}{\sigma^2}\boldsymbol{K}_m\right)\boldsymbol{B}^{-1}\right) \end{aligned}$$

$$+ \|\tau_{0m}\|_k^2 + \boldsymbol{\alpha}^\top \boldsymbol{K}_m(\hat{\boldsymbol{K}}_m^{-1}\boldsymbol{K}_m - \boldsymbol{I})\boldsymbol{\alpha} - n\Big\}. \tag{C.6}$$

Since the covariance function is bounded and continuous in $\boldsymbol{\theta}$, we have $\hat{\boldsymbol{K}}_m^{-1}\boldsymbol{K}_m - \boldsymbol{I} \to 0$ as $n \to \infty$. Hence, there exist positive constants c and ε such that for n large enough

$$\begin{aligned} -\log|\hat{\boldsymbol{K}}_m^{-1}\boldsymbol{K}_m| &< c, \\ \boldsymbol{\alpha}^\top \boldsymbol{K}_m(\hat{\boldsymbol{K}}_m^{-1}\boldsymbol{K}_m - \boldsymbol{I})\boldsymbol{\alpha} &< c, \\ \mathrm{tr}(\hat{\boldsymbol{K}}_m^{-1}\boldsymbol{K}_m\boldsymbol{B}^{-1}) &< \mathrm{tr}((\boldsymbol{I_n} + \varepsilon\boldsymbol{K}_m)\boldsymbol{B}^{-1}). \end{aligned} \tag{C.7}$$

Plugging (C.7) in (C.6), there exist positive constant δ, we have the inequality

$$\begin{aligned} &-\log p_{\sigma_0,\hat{\theta}}(\boldsymbol{y}_m|\boldsymbol{u}_m) + \log p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m},\boldsymbol{u}_m) \\ &\le \frac{1}{2}\Big\{2c + \log|\boldsymbol{B}| + \mathrm{tr}((\boldsymbol{I_n} + (\varepsilon + \sigma^{-2})\boldsymbol{K}_m)\boldsymbol{B}^{-1}) + \|\tau_{0m}\|_k^2 - n\Big\} \\ &\le c + \frac{1}{2}\log|\boldsymbol{B}| + \frac{1}{2}\|\tau_{0m}\|_k^2 + \delta. \end{aligned} \tag{C.8}$$

From the Representer Theorem (Lemma 2 in Seeger, Kakade and Foster (2008)), it shows

$$-\log p_{\sigma_0,\hat{\theta}}(\boldsymbol{y}_m|\boldsymbol{u}_m) + \log p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m},\boldsymbol{u}_m) \le c + \frac{1}{2}\log|\boldsymbol{B}| + \frac{1}{2}\|\tau_{0m}\|_k^2 + \delta$$

for all $\tau_{0m}(\cdot) \in \mathcal{H}$. The Lemma holds.

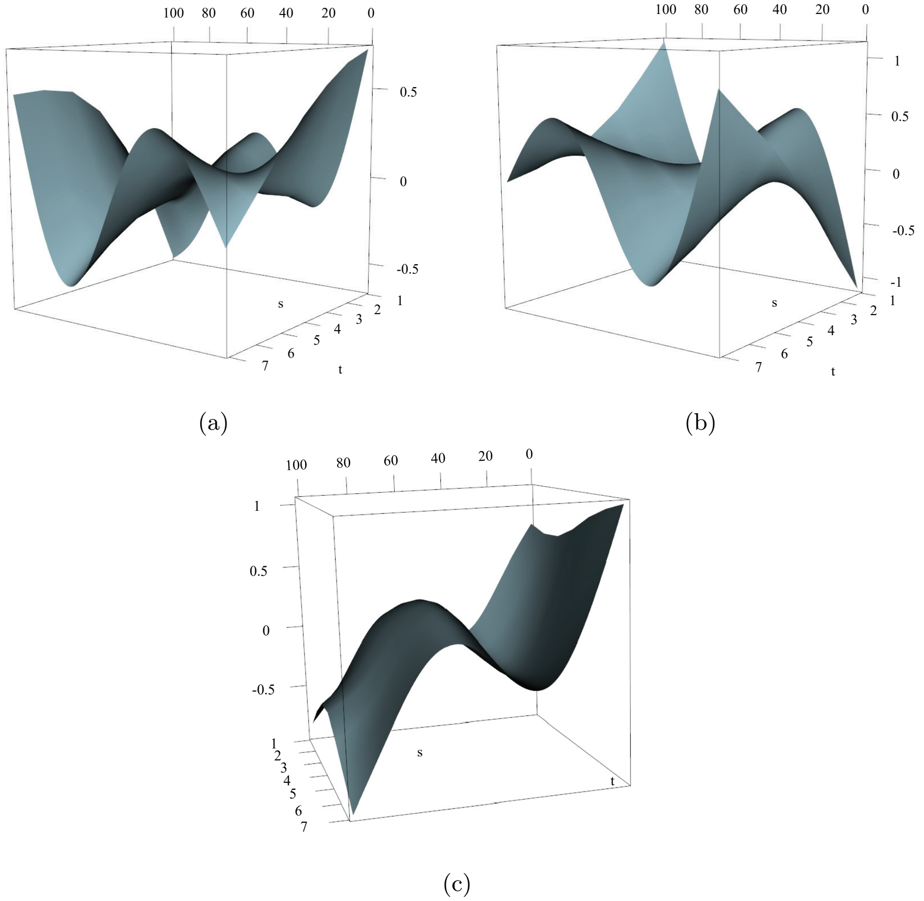
To prove Theorem 1, it needs
Condition (A): $\|\tau_{0m}\|_k$ is bounded and expected regret term

$$E_{\boldsymbol{u}_m}(\log|\boldsymbol{I} + \sigma^{-2}\boldsymbol{K}_m|) = o(n).$$

This condition, coming from Seeger, Kakade and Foster (2008), is called expected regret term. This term depends on covariance kernel function and covriates. For the kernel function k, we have spectrum decomposition

$$k(\boldsymbol{u},\boldsymbol{v}) = \sum_i \lambda_i\phi_i(\boldsymbol{u})\phi_i(\boldsymbol{v}),$$

where $\{\lambda_i\}$ and $\{\phi_i\}$ are eigenvalues and eigenvectors for k. If $\sum_i \lambda_i^2 < \infty$ which leads to λ_i decay rapidly to 0, then the expected regret term is bounded with $\sum_i \log(1 + c\lambda_s n)$. Hence, the condition (A) holds. More details can be found in



(a)

(b)

(c)

Figure 6. Estimations of functions $\beta_i(s,t)$, for $i = 1, 2, 3$. (a) 3D plot of $\beta_1(s,t)$; (b) 3D plot of $\beta_2(s,t)$; and (c) 3D plot of $\beta_4(s,t)$.

Section III in Seeger, Kakade and Foster (2008).

Proof of Theorem 1.

From definition of the information consistency, we can get that

$$
\begin{aligned}
&D[p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m},\boldsymbol{u}_m), p_{\sigma_0,\hat{\theta}}(\boldsymbol{y}_m|\boldsymbol{u}_m)]\\
&= \int_{\mathcal{F}} p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m},\boldsymbol{u}_m)\log\frac{p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m},\boldsymbol{u}_m)}{p_{\sigma_0,\hat{\theta}}(\boldsymbol{y}_m|\boldsymbol{u}_m)}d\boldsymbol{y}_m\\
&= \int_{\mathcal{F}} p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m},\boldsymbol{u}_m)\{-\log p_{\sigma_0,\hat{\theta}}(\boldsymbol{y}_m|\boldsymbol{u}_m) + \log p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m},\boldsymbol{u}_m)\}d\boldsymbol{y}_m.
\end{aligned}
$$

From Lemma 1, we obtain that

$$\begin{aligned}&\frac{1}{n}E_{\boldsymbol{u}_m}(D[p_{\sigma_0}(\boldsymbol{y}_m|\tau_{0m},\boldsymbol{u}_m),p_{\sigma_0,\hat{\theta}}(\boldsymbol{y}_m|\boldsymbol{u}_m)])\\&\leq\frac{c}{n}+\frac{1}{2n}E_{\boldsymbol{u}_m}(\log|\boldsymbol{I}+\sigma^{-2}\boldsymbol{K}_m|)+\frac{1}{2n}\|\tau_{0m}\|_k^2+\frac{\delta}{n},\end{aligned}$$

where c and δ are two positive constants. Since $\|\tau_{0m}\|_k$ is bounded and expected regret term $E_{\boldsymbol{u}_m}(\log|\boldsymbol{I}+\sigma^{-2}\boldsymbol{K}_m|)=o(n)$, Theorem 1 holds.

D. 3D Plots of Parameter Estimation for Movement Data

Estimations of functions $\beta_i(s,t),\ i=1,2,3$ are presented in Figure 6.

References

Cao, C., Shi, J. Q. and Lee, Y. (2018). Robust functional regression model for population-average and subject-specific inferences. *Statistical Methods in Medical Research* **27**, 3236–3254.

Cheng, Y., Shi, J. Q. and Eyre, J. (2017). Nonlinear mixed-effects scalar-on-function models and variable selection for Kinematic upper limb movement sata. *arXiv:1605.06779.*

Crambes, C. and Mas, A. (2013). Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli* **19**, 2627–2651.

Gervini, D. (2015). Dynamic retrospective regression for functional data. *Technometrics* **57**, 26–34.

Hastie, T. and Tibshirani, R. (1993). Varying coefficient models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **55**, 757–796.

Kim, K., Sentürk, D. and Li, R. (2011). Recent history functional linear models for sparse longitudinal data. *Journal of Statistical Planning and Inference* **141**, 1554–1566.

Kim, J. S., Staicu, A. M., Maity, A., Raymond, J., Carroll, R. J. and Ruppert, D. (2018). Additive function-on-function regression. *Journal of Computational and Graphical Statistics* **27**, 234–244.

Luo, R. and Qi, X. (2017). Function-on-function linear regression by signal compression. *Journal of the American Statistical Association* **112**, 690–705.

Malfait, N. and Ramsay, J. O. (2003). The historical functional linear model. *Canadian Journal of Statistics* **31**, 115–128.

Meyer, M. J., Coull, B. A., Versace, F., Cinciripini, P. and Morris, J. S. (2015). Bayesian function-on-function regression for multilevel functional data. *Biometrics* **71**, 563–574.

Müller, H. G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103**, 1534–1544.

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **53**, 539–572

Ramsay, J. O., Hooker, G. and Graves, S. (2010). *Functional Data Analysis with R and Matlab.* Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis.* 2nd Edition. Springer, New York.

Rao, J. N. (2003). *Small Area Estimation.* Wiley, New York.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning.* The MIT Press, Cambridge, Massachusetts.

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science* **6**, 15–32.

Scheipl, F., Staicu, A. M. and Greven, S. (2015). Functional Additive mixed models. *Journal of Computational and Graphical Statistics* **24**, 477–501.

Seeger, M. W., Kakade, S. M. and Foster, D. P. (2008). Information consistency of nonparametric Gaussian process methods. *IEEE Transactions on Information Theory* **54**, 2376–2382.

Serradilla, J., Shi, J., Cheng, Y., Morgan, G., Lambden, C. and Eyre, J. (2014). Automatic assessment of upper limb function during play of the action video game, circus challenge: validity and sensitivity to change. *Serious Games and Applications for Health (SeGAH)*, 1–7.

Shi, J., Cheng, Y., Serradilla, J., Morgan, G., Lambden, C., Gary, A., Christopher, P., Helen, R., Cassidy, T., Rochester, L. and Eyre, J. (2013). Evaluating functional ability of upper limbs after stroke using video game data. *Brain and Health Informatics* **8211**, 181–192.

Shi, J. and Choi, T. (2011). *Gaussian Process Regression Analysis for Functional Data.* Chapman and Hall, London.

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

Shi, J. Q., Wang, B., Will, E. J. and West, R. M. (2012). Mixed-effects GPFR models with application to dose-response curve prediction. *Statistics in Medicine* **31**, 3165–77.

Sun, X., Du, P., Wang, X. and Ma, P. (2018). Optimal penalized function-on-function regression under a reproducing kernel Hilbert space framework. *Journal of the American Statistical Association* **113**, 1601–1611.

Wang, B. and Shi, J. (2014). Generalized Gaussian process regression model for non-Gaussian functional data. *Journal of the American Statistical Association* **109**, 1123–1133.

Wang, Z., Shi, J. and Lee, Y. (2017). Extended T-process regression models. *Journal of Statistical Planning and Inference* **189**, 38–60.

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

Yao, F., Müller, H. G. and Wang, J. L. (2005b). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

Yuan, M. and Cai, T. (2010). A reproducing kernel Hilbert sapce approach to functional linear regression. *The Annals of Statistics* **38**, 3412–3444.

Zhanfeng Wang

Department of Statistics and Finance, The School of Management, University of Science and Technology of China, Hefei, China.

E-mail: zfw@ustc.edu.cn

Hao Ding

Department of Statistics and Finance,The School of Management, University of Science and

Technology of China, Hefei, China.

E-mail: dinghao@ustc.edu.cn

Zimu Chen

Department of Statistics and Finance, The School of Management, University of Science and Technology of China, Hefei, China.

E-mail: zmchen@mail.ustc.edu.cn

Jian Qing Shi

School of Mathematics and Statistics, Newcastle University, Newcastle, UK.

E-mail: j.q.shi@ncl.ac.uk

(Received July 2018; accepted February 2019)