Noname manuscript No.
(will be inserted by the editor)

Bagging-Enhanced Sampling Schedule for Functional Quadratic Regression

Hyungmin Rha · Ming-Hung Kao · Rong Pan

Received: date / Accepted: date

Abstract Establishing an optimal sampling schedule is a crucial step towards a precise inference of the underlying functional mechanism of a process, especially when data collection is expensive/difficult. This work is concerned with optimal sampling plans for predicting a scalar response using a functional predictor when a quadratic regression relationship is present. An optimality criterion for selecting the best sampling schedules is derived, and some important properties of the criterion are provided. In addition, a bootstrap aggregating (bagging) strategy is proposed to enhance the quality of the obtained sampling schedule.

Keywords functional data analysis · functional principal component · functional regression model · bagging

Hyungmin Rha
School of Mathematical and Statistical Sciences, Arizona State University

Ming-Hung Kao
School of Mathematical and Statistical Sciences, Arizona State University
Tel.: +1 (480) 965-3466
E-mail: mkao3@asu.edu

Rong Pan
Ira A. Fulton Schools of Engineering, Arizona State University

1 Introduction

Functional data analysis is a contemporary statistical methodology that has a wide range of applications such as brain imaging [11, 22], medical research [23], pharmaceutics [13, 25], and econometrics [16, 19]. However, having a precise analysis of functional data can be challenging. This is partly because the true underlying function is defined over a continuous domain, but the observations contaminated by measurement errors are almost always collected over a discretized domain. In some cases, only very few measurements can be acquired to study the underlying function due to limited resources (e.g., time and/or money), and other practical reasons. For such a situation, a judiciously selected sampling plan for collecting informative data to yield a precise analysis result is crucial. There have been several studies on this direction. For example, Li [14], and Li and Xiao [15] discussed the design issues for classification of functional data. Wu *et al.* [26] considered the D-optimal design for capturing the between-subject variability of the underlying trajectories. Ji and Müller [10], and Park *et al.* [17] obtained optimal sampling schedules for functional linear regression to yield a high precision in recovering the predictor function, and in predicting a scalar response. Rha *et al.* [21] put forward a probabilistic subset search (PSS) algorithm to identify optimal sampling schedules for both scalar-on-function, and function-on-function regression models.

The previous works on selecting designs for functional regression mainly focused on models that only allow a 'simple linear' relationship between the response and the predictor function. Throughout this paper, we refer to these models as functional simple linear regression (FSLR) models; see also [20]. In many real applications, a complex association between the response and the predictor may exist, and the FSLR model can underfit the data, resulting in, e.g., an inaccurate prediction of the response. To address this issue,

Yao and Müller [29] extended the FSLR models to the functional quadratic regression (FQR) models. The FQR is demonstrated to be useful. However, to our knowledge, the selection of optimal sampling schedules for this rather complex setting has not been investigated so far. To address this lack, our first contribution here is on extending the optimal design methods to FQR models. We derive an optimality criterion for comparing sampling schedules under the FQR model, and present some important properties of this criterion. This new criterion contains the one for FSLR as a special case, and can be considered for both types of models.

However, the optimality criterion for functional regression depends on several unknown parameters of the model, such as the pairs of the eigenfunction and eigenvalue, and the variances and covariances of the noisy observations of the predictor function, $X(t)$. With this issue, the previously mentioned studies considered a locally optimal design approach similar to that of [7]. Specifically, the unknown parameters are replaced by their estimates from, e.g., a pilot study. The resulting locally optimal designs thus rely much on one single estimation of the parameters. These designs may perform well (or is optimal) when all the parameter estimates are close to (or the same as) the true values. But, they unfortunately can suffer an efficiency loss, especially when a large pilot study for rendering reliable parameter estimates is unavailable (e.g., due to limited resources). To alleviate this design dependence problem, we propose here an easy-to-implement strategy that combines the PSS algorithm with the bagging (bootstrap aggregating) technique to obtain bagging-enhanced sampling schedule (BESS) designs for functional regression. Through simulations, we show that, without much computing time, the BESS approach can generate designs that outperform locally optimal designs. A significant improvement is achieved by the BESS designs when the unknown parameters are estimated from small- to medium-sized pilot data. For cases where locally optimal de-

signs perform well, the BESS designs still attain a comparable or slightly better performance.

This paper is organized as follows. In Section 2, we introduce our notation and the functional regression models that we consider. We also derive the optimality criterion for evaluating competing designs for the FQR model, and provide some important properties of the optimality criterion. We then describe the BESS approach in Section 3, and present some simulation studies in Section 4. Some real applications can be found in Section 5, and a discussion is in Section 6.

2 Model and Optimality Criterion

Let $X(\cdot)$ be a square integrable stochastic process defined over a continuous compact domain $\mathcal{T}$ with mean function $\mathrm{E}\{X(t)\} = \mu_X(t)$ and covariance function $\mathrm{Cov}\{X(s), X(t)\} = \Gamma(s,t)$. By Mercer's Theorem, $\Gamma(s,t) = \sum_{k=1}^{\infty} \rho_k \psi_k(s)\psi_k(t)$, where $\psi_k(\cdot)$ and ρ_k, with $\rho_1 \geq \rho_2 \geq ... \geq 0$, are the k^{th} eigenfunction and eigenvalue of the covariance operator of $X(t)$, respectively. With some mild conditions [8], $X(t)$ can be decomposed as $X(t) = \mu_X(t) + \sum_{k=1}^{\infty} \zeta_k \psi_k(t)$, where the functional principal component (FPC) scores, $\zeta_k = \int_{\mathcal{T}} \{X(t) - \mu_X(t)\}\psi_k(t)dt, k = 1, 2, ...$, are uncorrelated with mean 0 and variance ρ_k. We denote the noisy observation of $X(t)$ at $t = t_j$ as $U_j = X(t_j) + \varepsilon_j$, where $t_j \in \mathcal{T}$ and ε_j's are independent random noise with mean 0 and variance σ_X^2, $j = 1, ..., L$. In addition, we assume that ζ_k and ε_j are independent and normally distributed. We also consider a scalar response Y paired with the functional predictor $X(t)$, and the following FQR model [29],

$$E(Y|X) = \alpha + \int_{\mathcal{T}} \beta(t)X^c(t)dt + \int_{\mathcal{T}}\int_{\mathcal{T}} \gamma(s,t)X^c(s)X^c(t)dsdt, \qquad (1)$$

where $X^c(t) = X(t) - \mu_X(t)$, α is an unknown intercept, and $\beta(t)$ and $\gamma(s,t)$ are square integrable coefficient functions for the linear and quadratic terms, respectively. Since $\psi_k(\cdot)$'s form a basis of L^2-space on a compact continuous domain $\mathcal{T}$, we write

$$\beta(t) = \sum_{k=1}^{\infty} \beta_k \psi_k(t); \quad \gamma(s,t) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \gamma_{kl}\psi_k(s)\psi_l(t), \tag{2}$$

where $\sum_{k=1}^{\infty} \beta_k^2 < \infty$ and $\sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \gamma_{kl}^2 < \infty$. With (1) and (2), and the orthonormality of eigenfunctions, we then have

$$E(Y|X) = \alpha + \sum_{k=1}^{K} \beta_k \zeta_k + \sum_{k=1}^{K}\sum_{l=1}^{K} \gamma_{kl}\zeta_k\zeta_l, \tag{3}$$

where K is possibly infinite. In our case studies, we choose K so that 99% of variability of $X(t)$ is captured. For a new subject with p observations, we will observe $\boldsymbol{U} = (U_1, ..., U_p)^T$ at $\mathbf{t} = \{t_1, ..., t_p\}$, $t_1 < \ldots < t_p$. A prediction of Y for this subject is

$$E(E(Y|X)|\boldsymbol{U}) = \alpha + \sum_{k=1}^{K} \beta_k \tilde{\zeta}_k + \sum_{k=1}^{K}\sum_{l=1}^{K} \gamma_{kl}\widetilde{\zeta_k\zeta_l}, \tag{4}$$

where $\tilde{\zeta}_k = E(\zeta_k|\boldsymbol{U})$ and $\widetilde{\zeta_k\zeta_l} = E(\zeta_k\zeta_l|\boldsymbol{U})$. As in [30], the best linear unbiased predictor for ζ_k is $\tilde{\zeta}_k = \rho_k \psi_k^T(\mathbf{t}) \Gamma_*^{-1}(\mathbf{t})(\boldsymbol{U} - \mu_X(\mathbf{t}))$, where $\psi_k(\mathbf{t}) = (\psi_k(t_1), ..., \psi_k(t_p))^T$, $\Gamma_*(\mathbf{t}) = \mathrm{Cov}(\boldsymbol{U})$, and $\mu_X(\mathbf{t}) = (\mu_X(t_1), ..., \mu_X(t_p))^T$. For simplicity, we omit $\mathbf{t}$ from $\Gamma_*(\mathbf{t})$, $\psi_k(\mathbf{t})$ and $\mu_X(\mathbf{t})$, and denote them as Γ_*, $\boldsymbol{\psi}_k$ and $\boldsymbol{\mu}_X$, respectively. Our aim is to find the sampling schedule $\mathbf{t}^*$ that minimizes the mean squared error for predicting Y; i.e.,

$$\begin{aligned} E\Big(E(Y|X) - E(E(Y|X)|\boldsymbol{U})\Big)^2 &= \mathrm{Var}\Big(E(Y|X) - E(E(Y|X)|\boldsymbol{U})\Big) \\ &= \mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T R + 2(\mathrm{G}R)^2\big] - \mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T H\Gamma_*^{-1}H^T + 2(\mathrm{G}H\Gamma_*^{-1}H^T)^2\big] \geq 0, \end{aligned} \tag{5}$$

where R is a $K \times K$ diagonal matrix whose k^{th} diagonal element is ρ_k, $\boldsymbol{b} = (\beta_1, ..., \beta_K)^T$, G is a $K \times K$ matrix whose $(i,j)^{th}$ element is γ_{ij}, and $H = R\Psi^T$ with $\Psi = (\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_K)$. The detailed derivation of (5) can be found in Appendix A. Since $\mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T R + 2(\mathrm{G}R)^2\big]$ does not depend on the sampling schedule, the larger-the-better criterion for finding optimal $\mathbf{t}^*$ can be set to:

$$F_Y(\mathbf{t}) = \mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T H \Gamma_*^{-1} H^T + 2(\mathrm{G}H\Gamma_*^{-1}H^T)^2\big]. \tag{6}$$

We note that the criterion $F_Y(\mathbf{t})$ in (6) is reduced to the optimality criterion for the FSLR model when G is zero; see, e.g., [17]. But when the quadratic relationship is present, $F_Y(\mathbf{t})$ includes an additional term $\mathrm{tr}\big[2(\mathrm{G}H\Gamma_*^{-1}H^T)^2\big]$. With this observation, the approach that we proposed here for FQR models can thus be easily applied to cases with FSLR models. In addition, we see from (5) that $\mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T R + 2(\mathrm{G}R)^2\big]$ gives an upper bound of $F_Y(\mathbf{t})$. We thus can define the relative efficiency of a design as follows:

$$RE_Y(\mathbf{t}) = \frac{F_Y(\mathbf{t})}{\mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T R + 2(\mathrm{G}R)^2\big]}. \tag{7}$$

Following [17], we now provide some useful properties of our optimality criterion in the next two theorems. Their proofs can be found in Appendix B.

Theorem 1 *Suppose* $\mathbf{t} \subseteq \tilde{\mathbf{t}}$ *where* $\mathbf{t}$ *and* $\tilde{\mathbf{t}}$ *are* p*- and* $(p+c)$*-point designs (sampling schedules), respectively, for some positive integer* c*. Then,* $F_Y(\mathbf{t}) \leq F_Y(\tilde{\mathbf{t}})$.

Theorem 1 supports the statement that acquiring additional observations from $X(t)$ can improve the response prediction precision. The next theorem suggests that $F_Y(\mathbf{t})$ approaches to its theoretical upper bound when the number of observations p goes to infinity.

Theorem 2 *Let* $\mathbf{t}_p = \{0, 1/p, ..., (p-1)/p\}$. *Then*

$$\lim_{p\to\infty} F_Y(\mathbf{t}_p) = \mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T R + 2(\mathrm{G}R)^2\big].$$

In practice, the unknown parameters involved in the optimality criteria (6) and (7) need to be estimated from, e.g., a pilot study. There exist several discussions on estimating these parameters from data [9,18,27,30,31,29]. For demonstration purposes, we use the `FPCA` function of `'fdapace'` package [6] in `R` to estimate the matrices R, H and Γ_*, and the `FPCQuadReg` function of `'PACE'` package [28] in `MATLAB`, which is converted into `R` in our implementation, to estimate the vector $\boldsymbol{b}$ and the matrix G of coefficients. Some other estimation methods can be considered. The parameter estimates give an estimated optimality criterion, $\hat{F}_Y$, and an estimated relative efficiency $\widehat{RE}_Y$. Similarly to the previous works, one may then consider a discretized domain of $\mathcal{T}$ that has N ($> p$) equally-spaced points, and select an optimal p-point sampling schedule out of the N points to maximize $\hat{F}_Y$ (or $\widehat{RE}_Y$).

To find an optimal design, an approach considered by Park *et al.* [17] is the exhaustive search algorithm that evaluates all the $\binom{N}{p}$ designs. This algorithm guarantees the best design within the discretized domain, but is computationally expensive. It can easily become infeasible for realistic problems. Ji and Müller [10] suggested a greedy search algorithm that sequentially adds the time point yielding the greatest improvement in the value of the optimality criterion at each iteration. However, the greedy search can sometimes be trapped in a poor local solution. Recently, Rha *et al.* [21] proposed the Probabilistic Subset Search (PSS) algorithm for generating (nearly-)optimal designs for FSLR models. Here, we adapt the PSS algorithm with its default settings to obtain sampling schedule for the FQR model. A pseudo code for this algorithm is provided in Appendix C.

A major drawback of the previously mentioned approaches is that the quality of the obtained design depends heavily on the quality of the estimated criterion $\hat{F}_Y$, and it does not take the estimation uncertainty into account. Consequently, the obtained designs might not be as efficient, especially when the prior information about the unknown parameters in F_Y is vague, and/or the pilot study for estimating these parameters does not have a large sample size. To address this issue, we consider to enhance the quality of the obtained designs by incorporating a bagging strategy into our design approach. To the best of our knowledge, this strategy has not previously been utilized in optimal design problems as the one we consider here. In the next section, we introduce our proposed Bagging-Enhanced Sampling Schedule (BESS) approach for obtaining optimal designs for functional regression models.

3 Bagging-Enhanced Sampling Schedule

As previously described, our aim is at an optimal sampling plan that maximizes the optimality criterion F_Y. Since F_Y involves unknown parameters, we follow previous works to consider a surrogate criterion, namely the estimated $\hat{F}_Y$. The obtained design can then be viewed as a locally optimal design [7]. Clearly, the quality of the estimated $\hat{F}_Y$ is crucial to the 'true' performance of a locally optimal sampling schedule. However, it is not always possible to obtain a good estimate, and as a result, the sampling schedule $\hat{\mathbf{t}}$ obtained by maximizing the surrogate $\hat{F}_Y(\mathbf{t})$ may not render data that are as informative as we expect.

Here, we propose an approach that borrows the strengths of the bagging technique to enhance the quality of the obtained sampling schedule. Bagging is known as a powerful computational approach for improving unstable estimations [2,3]. We utilize this strategy to give an improved estimate of F_Y to generate high quality designs. The detailed steps of our proposed BESS ap-

proach are provided below. There, we consider situations where a pilot data set of n subjects is available for estimating the unknown parameters in F_Y of (6).

Step 1. From the pilot data, generate a bootstrap sample of n subjects through subject-wise re-sampling with replacement; repeat this step B times to obtain B bootstrap samples.

Step 2. With the b^{th} bootstrap sample in Step 1, obtain estimates of the unknown parameters in $F_Y(\mathbf{t})$ to give $\hat{F}_{Y,b}(\mathbf{t})$, an estimate of $F_Y(\mathbf{t})$ for the given design $\mathbf{t}$; $b = 1, ..., B$.

Step 3. Find $\hat{\mathbf{t}}^B$ that maximizes

$$\hat{F}_Y^B(\mathbf{t}) = \frac{1}{B}\sum_{b=1}^{B} \hat{F}_{Y,b}(\mathbf{t}). \tag{8}$$

In contrast to the previously proposed methods, the BESS approach does not fully rely on the single $\hat{F}_Y(\mathbf{t})$ estimated from the pilot data set. It instead uses the 'bagging estimate' $\hat{F}_Y^B$ of F_Y as the surrogate criterion for selecting optimal sampling schedule. To search for the optimal schedule $\hat{\mathbf{t}}^B$ in Step 3 of the BESS approach, we adapt the PSS algorithm of [21], which was developed for obtaining locally optimal designs for FSLR models. We note that the BESS approach can be easily applied to some other situations, such as recovering the trajectory of $X(t)$, and predicting a functional or scalar response in FSLR as considered in the previous works [10,17,21]. To demonstrate the applicability of our proposed approach, we present below some simulation studies on obtaining optimal sampling schedules for having a precise recovery of the predictor function $X(t)$, and a precise prediction of the response Y. Additional simulation results can also be found in the supplementary material.

4 Simulation Study

For demonstration purposes, we consider here two types of eigenfunctions for the predictor function $X(t)$, which are polynomial basis functions and Fourier basis functions. We assume that there are three eigenfunctions paired with non-zero eigenvalues. Results for two other types of eigenfunctions are presented in the supplementary material, and they convey similar information on demonstrating the usefulness of our proposed approach. Without loss of generality, we assume that $\mu_X(t) = 0$ for all $t \in \mathcal{T} = [0, 1]$ and $\alpha = 0$. The measurement errors are assumed to follow the standard Normal distribution.

For the first scenario, we set the three eigenfunctions of $X(t)$ to

$$\psi_1(t) = \sqrt{3}(1 - 2t), \quad \psi_2(t) = \sqrt{5}(1 - 6t + 6t^2), \quad \text{and}$$
$$\psi_3(t) = \sqrt{7}(1 - 12t + 30t^2 - 20t^3),$$

and the eigenvalues are $\rho_k = 5, 3$, and 1, respectively. The $\boldsymbol{b}$ vector and G matrix are assumed to be

$$\boldsymbol{b} = (1, -0.5, 0.5)^T \quad \text{and} \quad \mathrm{G} = \begin{bmatrix} 0.20 & -0.05 & 0.25 \\ -0.05 & 0.35 & -0.05 \\ 0.25 & -0.05 & 0.40 \end{bmatrix}.$$

For the second scenario, the eigenfunctions of $X(t)$ are

$$\psi_1(t) = \sqrt{2}\sin(2\pi t), \quad \psi_2(t) = \sqrt{2}\cos(4\pi t), \quad \text{and} \quad \psi_3(t) = \sqrt{2}\sin(6\pi t),$$

and the corresponding eigenvalues are $\rho_k = 8, 2, 1$, respectively. In addition,

$$\boldsymbol{b} = (1, -0.5, -1)^T \quad \text{and} \quad \mathrm{G} = \begin{bmatrix} 0.25 & -0.05 & 0.10 \\ -0.05 & 0.30 & -0.02 \\ 0.10 & -0.02 & 0.30 \end{bmatrix}.$$

We also write $\mathrm{Var}\big(E(Y|X)\big) = \mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T R + 2(\mathrm{G}R)^2\big] = w_l \mathrm{tr}(\boldsymbol{b}\boldsymbol{b}^T R) + w_q \mathrm{tr}[2(\mathrm{G}R)^2]$ with $(w_l, w_q) = (1, 1)$. The values of $\mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T R\big]$ and $\mathrm{tr}\big[2(\mathrm{G}R)^2\big]$ reflect the proportion of variability due to the linear term and the quadratic term, respectively, in the FQR model, and w_l and w_q can be viewed as the corresponding weights. To study the effects of these two components on the obtained designs, we vary $\tau = w_q/(w_l + w_q)$, i.e., the 'relative importance' of the quadratic term, by considering different values for (w_l, w_q).

In the following simulation studies, we first use the PSS algorithm to search for a design maximizing the true F_Y of each scenario described previously. We demonstrate that the obtained 'true' optimal design $\mathbf{t}^*$ outperforms random design, equally-spaced design, and the optimal design obtained under the FSLR model when the quadratic term is ignored. We then obtain the locally optimal designs, $\hat{\mathbf{t}}$, by optimizing the $\hat{F}_Y$ estimated from simulated pilot data, and the BESS designs, $\hat{\mathbf{t}}^B$, by the BESS approach described in Section 3. We show that $\hat{\mathbf{t}}^B$ outperforms $\hat{\mathbf{t}}$ in the sense that it reduces the absolute relative error (ARE) to be defined later. We also investigate the performance of designs obtained under a situation where the Gaussian assumption is violated.

4.1 Designs optimizing the true F_Y

Here, we compare the performance of the design that maximizes F_Y with random designs, equally-spaced designs, and the optimal designs obtained in [21] for the FSLR model. Our designs are obtained with the PSS algorithm over the discretized $\mathcal{T} = [0, 1]$ with grid size 0.01. With the simulation settings described above, the FPC score ζ_k of $X(t)$ is assumed to follow a Normal distribution with mean 0 and variance ρ_k, for $k = 1, 2, 3$. For this comparison, we generate $1,000$ 'true' responses Y by using Eq.(3) under the settings for each scenario. For each true response, we then generate 100 sets of p noisy

observations $\boldsymbol{U}$ for each p-point design being compared; $p = 3, ..., 7$. Each set of p observations is then used to predict Y with $E(E(Y|X)|\boldsymbol{U})$ as introduced in Section 2. We then calculate the root mean squared error (RMSE) of the 100 predicted responses for each true response. This gives 1000×100 RMSEs for each design, and the average of these RMSEs is compared.

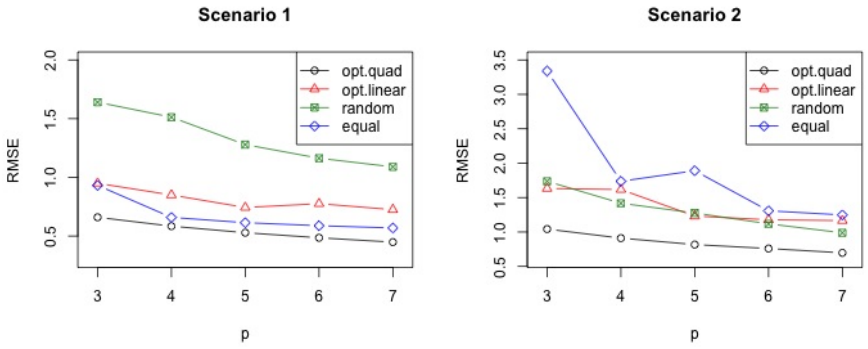

Fig. 1 Average RMSEs with $\tau = 50\%$; opt.quad and opt.linear represent the designs for FQR and FSLR models, respectively.

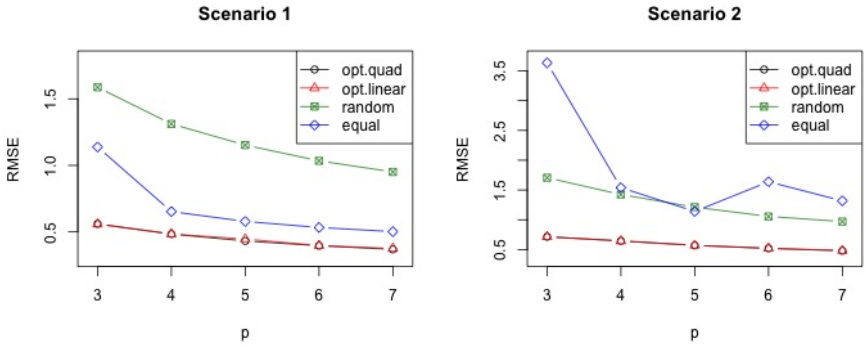

Fig. 2 Average RMSEs with $\tau = 1\%$. opt.quad and opt.linear represent the designs for FQR and FSLR models, respectively.

As shown in Figure 1 where we set $(w_l, w_q) = (1, 1)$, and $\tau = 50\%$, the optimal design for the FQR model clearly outperforms the other designs in

minimizing the RMSE. In Figure 2, we set $(w_l, w_q) = (2, 0.02)$ to give a small $\tau = 1\%$. As expected, the optimal sampling schedules for FSLR and those for FQR perform similarly.

4.2 BESS vs. locally optimal designs

We now consider a more practical situation where pilot data are used to give the estimated criterion $\hat{F}_Y$. We compare the BESS design $\hat{\mathbf{t}}^B$, which maximizes $\hat{F}_Y^B$ in (8), and the locally optimal design $\hat{\mathbf{t}}$ that optimizes $\hat{F}_Y$. For each scenario, we simulate 200 pilot data sets with $U_{ij} = X_i(t_j) + \varepsilon_{ij}$ as the j^{th} observation of the i^{th} subject; $j = 1, ..., J_i$, $i = 1, ..., n$. We consider pilot data sets of different sizes, including $n = 25$, 50, and 100. The number of observations for the i^{th} subject in a data set is randomly generated with $J_i \sim \text{Uniform}\{2, ..., 10\}$. The random errors ε_{ij}'s are i.i.d. $N(0, 1)$. For each simulated data set, we obtain $\hat{F}_Y$ and $\hat{F}_Y^B$ with $B = 50$, and use the PSS algorithm to obtain $\hat{\mathbf{t}}$ and $\hat{\mathbf{t}}^B$ that maximize $\hat{F}_Y$ and $\hat{F}_Y^B$, respectively, with $p = 3, \ldots, 7$. In addition, we demonstrate the applicability of the BESS approach in finding optimal designs for having a precise recovery of the predictor function $X(t)$, another optimal design issue of interest for functional regression. Following [21], we set the optimality criterion for recovering $X(t)$ to $F_X(\mathbf{t}) = \text{tr}(H\Gamma_*^{-1}H^T)$, and aim at a design maximizing this criterion.

For comparison purposes, we consider the ARE of a design $\mathbf{t}$ to the 'true' optimal design $\mathbf{t}^*$ that is obtained by using the PSS algorithm to maximize the true criterion F. Depending on the study objective (predicting Y or recovering $X(t)$), F can be F_Y or F_X.

$$ARE(\mathbf{t}) = \frac{|\, F(\mathbf{t}) - F(\mathbf{t}^*) \,|}{F(\mathbf{t}^*)}.$$

Table 1 Average (standard deviation) of 200 *ARE*s (×100%) for predicting the scalar response Y under Scenarios 1 and 2; $\hat{\mathbf{t}}$ and $\hat{\mathbf{t}}^B$ represent the locally optimal, and BESS designs, respectively.

		Scenario 1		*Scenario 2*	
		$\hat{\mathbf{t}}$	$\hat{\mathbf{t}}^B$	$\hat{\mathbf{t}}$	$\hat{\mathbf{t}}^B$
$p=3$	$n=25$	6.576 (10.37)	3.381 (6.008)	12.26 (12.07)	11.97 (10.03)
	$n=50$	3.605 (6.544)	2.254 (5.043)	9.256 (6.512)	8.616 (5.551)
	$n=100$	1.362 (1.257)	1.024 (1.129)	7.198 (5.061)	6.843 (4.595)
$p=4$	$n=25$	5.736 (9.799)	1.542 (2.402)	10.84 (11.01)	9.764 (6.376)
	$n=50$	1.857 (4.219)	1.048 (1.164)	6.617 (5.226)	5.318 (3.958)
	$n=100$	0.826 (1.332)	0.771 (1.071)	6.161 (4.641)	5.016 (3.544)
$p=5$	$n=25$	4.094 (7.141)	1.094 (0.968)	8.614 (8.097)	6.599 (5.477)
	$n=50$	1.330 (2.124)	0.848 (0.874)	5.713 (5.219)	4.105 (3.066)
	$n=100$	0.543 (1.377)	0.492 (0.555)	4.465 (3.846)	3.522 (3.242)
$p=6$	$n=25$	2.952 (4.769)	1.069 (0.583)	6.913 (6.882)	4.709 (3.284)
	$n=50$	1.242 (1.965)	0.820 (0.758)	4.725 (4.157)	3.315 (2.631)
	$n=100$	0.480 (0.866)	0.437 (0.348)	3.526 (2.946)	2.792 (2.314)
$p=7$	$n=25$	3.681 (6.783)	1.156 (0.534)	5.588 (5.780)	4.012 (2.609)
	$n=50$	1.069 (2.088)	0.825 (0.468)	4.154 (3.676)	2.832 (1.769)
	$n=100$	0.190 (0.398)	0.283 (0.300)	3.229 (2.668)	2.343 (1.728)

A comparison of the ARE of the BESS designs $\hat{\mathbf{t}}^B$ with that of the locally optimal designs $\hat{\mathbf{t}}$ is presented in Tables 1 and 2. We consider the prediction of Y in Table 1, and the recovery of $X(t)$ in Table 2. For most cases, the BESS method significantly reduces both the average ARE, and the uncertainty of the achieved design performance; the latter is reflected in the much smaller standard deviation of the AREs of the BESS designs than that of the locally optimal designs. A greater improvement is seen in the cases with a smaller sample size n. This might be because of the improved estimation quality of $\hat{F}_Y$ and $\hat{F}_X$ with an increased sample size. Nevertheless, we still observe obvious improvements in many cases even when the sample size is as large as $n = 100$. In addition, as the number of subjects increases, ARE decreases across all cases, which again might be due to an improved quality of estimation.

Table 2 Average (standard deviation) of 200 *ARE*s (×100%) for recovering the predictor function $X(t)$ under Scenarios 1 and 2; $\hat{\mathbf{t}}$ and $\hat{\mathbf{t}}^B$ represent the locally optimal, and BESS designs, respectively.

		Scenario 1		*Scenario 2*	
		$\hat{\mathbf{t}}$	$\hat{\mathbf{t}}^B$	$\hat{\mathbf{t}}$	$\hat{\mathbf{t}}^B$
$p=3$	$n=25$	4.223 (5.880)	2.840 (4.606)	8.247 (6.477)	7.248 (5.526)
	$n=50$	2.424 (3.887)	1.792 (2.793)	5.475 (4.714)	4.827 (3.960)
	$n=100$	0.941 (1.941)	0.806 (1.536)	4.134 (3.328)	4.080 (3.381)
$p=4$	$n=25$	3.630 (5.224)	1.808 (2.445)	6.641 (5.696)	4.885 (4.304)
	$n=50$	1.249 (2.741)	0.886 (1.220)	4.823 (3.736)	3.331 (2.941)
	$n=100$	0.607 (1.187)	0.526 (0.602)	3.798 (2.776)	3.384 (2.445)
$p=5$	$n=25$	2.590 (4.198)	0.484 (0.987)	5.637 (5.081)	3.392 (3.291)
	$n=50$	1.287 (2.309)	0.514 (0.634)	4.003 (3.346)	2.569 (2.323)
	$n=100$	0.487 (1.185)	0.282 (0.402)	2.834 (2.616)	2.042 (1.822)
$p=6$	$n=25$	1.991 (3.449)	0.783 (0.962)	4.700 (4.374)	3.016 (3.196)
	$n=50$	1.043 (1.770)	0.569 (0.476)	3.344 (2.931)	2.284 (1.937)
	$n=100$	0.665 (1.374)	0.468 (0.431)	2.744 (2.183)	2.053 (1.554)
$p=7$	$n=25$	1.747 (2.880)	0.618 (0.423)	4.430 (4.509)	2.723 (2.498)
	$n=50$	0.852 (1.515)	0.461 (0.393)	2.761 (2.415)	2.141 (1.900)
	$n=100$	0.519 (1.301)	0.383 (0.371)	2.610 (2.170)	1.976 (1.598)

4.3 Robustness to Gaussian assumption

The optimality criterion in (6) is derived under the Gaussian assumption of the FPC scores. In this subsection, we show that the designs obtained from our method consistently outperform other designs even when the Gaussian assumption of the FPC scores is violated. To demonstrate the robustness of our method to a violation of the Gaussian assumption, we assume that the k^{th} FPC score ζ_k, $k=1,2,3$ now follows a mixture of two Gaussian distributions, namely $N(-\sqrt{1.8\rho_k}, 0.1\rho_k)$ with probability 1/3, and $N(\sqrt{0.45\rho_k}, 0.1\rho_k)$ with probability 2/3. This mixture distribution has mean 0 and variance ρ_k, and is asymmetric and bimodal. We then follow the same comparison procedure as in Subsection 4.1 to obtain Table 3; we again consider $(w_l, w_q) = (1,1)$, and

Table 3 Average RMSEs under a violation of Gaussian assumption. opt.quad and opt.linear represent the designs for FQR and FSLR, respectively.

		$p=3$	$p=4$	$p=5$	$p=6$	$p=7$
Scenario 1	opt.quad	0.6762	0.5896	0.5309	0.4923	0.4541
	opt.lin	0.9936	0.9035	0.7879	0.6962	0.7746
	random	1.6660	1.4719	1.3465	1.1889	1.0967
	equal	0.9754	0.6648	0.6204	0.5924	0.5744
Scenario 2	opt.quad	1.0825	0.9379	0.8546	0.7863	0.7219
	opt.lin	1.6229	1.5910	1.1927	1.1423	1.1241
	random	1.6994	1.4847	1.3000	1.1456	1.0396
	equal	3.2124	1.8304	1.9360	1.4593	1.3358

thus $\tau = 50\%$. It can be seen that the results in Table 3 render consistent information as Figure 1, even when the Gaussian assumption is violated.

5 Application

In this section, we consider two real data sets, including the Alzheimer's disease neuroimaging data, and the Berkeley growth data set. Treating these as pilot data sets, we apply our proposed method to find a sampling schedule for predicting the scalar response of interest in each of these two applications.

5.1 Alzheimer's Disease Neuroimaging Data

The data used here are from the data sheet "148-n3355-FDG-ALL-Info" in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database for Alzheimer's disease (AD). The AD is a progressive, degenerative brain disorder that is currently known to be irreversible. For this data set, we treat the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-cog) collected over some years as the predictor function $X(t)$, and the hypometabolic convergence index (HCI) score of each AD patient obtained from her/his last magnetic resonance imaging (MRI) scan as the scalar response Y. The HCI is a single

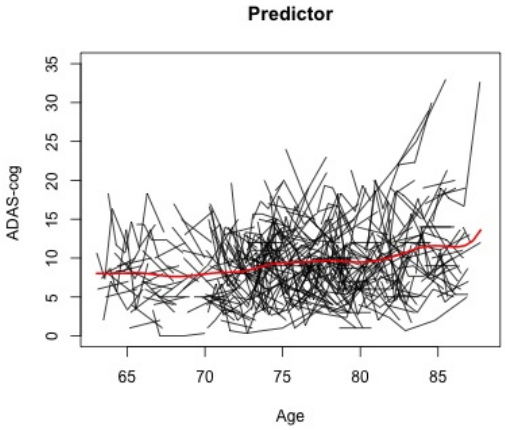

Fig. 3 Alzheimer's disease assessment scale-cognitive subscale of 269 patients. Red thick curve represents the mean function.

metric that measures the extent to which the pattern and magnitude of cerebral hypometabolism in the fluorodeoxyglucose positron emission tomography (FDG-PET) image correspond to ADAS-cog of a patient. The ADAS-cog is a clinical rating for evaluating the level of cognitive impairment in AD patients, and consists of 11 tasks to measure cognitive abilities that are often referred to as the core symptoms of AD; see, e.g., [12].

Most of the measurements in the data set are collected between age 63 and 88. We thus focus on the data obtained from patients over this compact domain $\mathcal{T}$. A total of 269 patients have ADAS-cog ratings within this age range, and the number of measurements of each patient varies from 1 to 8 with a median of 2; see Figure 3 for the Spaghetti plot of the predictor functions over $\mathcal{T}$ of these patients, and their mean function. All these 269 patients have the corresponding HCI scores.

For the previously described data, we fit both the FSLR and FQR models. Following [29], we calculate the corresponding quasi-R^2 for the two models to demonstrate that the FQR model provides a better fit to the data than

Table 4 Optimal sampling schedules and relative efficiencies for predicting HCI through ADAS-cog curve when the number of observations $p = 2, ..., 6$.

	Optimal Sampling Schedule	Relative Efficiency
$p = 2$	69.5, 87.0	0.6221
$p = 3$	63.0, 69.0, 88.0	0.7779
$p = 4$	63.0, 69.0, 69.5, 88.0	0.8171
$p = 5$	63.0, 69.0, 69.5, 87.5, 88.0	0.8522
$p = 6$	63.0, 69.0, 69.5, 70.0, 87.5, 88.0	0.8685

the FSLR model. As noted in [29], the quasi-R^2 can give a straightforward model comparison, and its value is not automatically improved by including additional predictors in the model. This is in contrast to the R^2 for the ordinary linear regression (with scalar predictors). We present below the formula of the quasi-R^2.

$$\hat{R}^2_M = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i^M)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}. \tag{9}$$

Here, the subscript M of $\hat{R}^2$ is used to indicate the model that is being evaluated. In particular, for the FSLR model, we set $M = L$; and $M = Q$ corresponds to the FQR model. Similarly, $\hat{Y}_i^M$ is the predicted value of Y under Model M. Specifically, $\hat{Y}^Q$ is obtained with the method described in Section 2; see also [29]. With the FSLR model, $\hat{Y}^L = \hat{\alpha} + \sum_{k=1}^K \hat{\beta}_k \hat{\zeta}_k$ as described in, e.g., [21] and the references therein; K is defined as in Section 2, and is selected so that 99% of the variability of $X(t)$ is captured. For this data set, we have $\hat{R}^2_L = 0.0694$, and $\hat{R}^2_Q = 0.2139$; the latter model has a significantly improved value for the quasi-R^2.

With the FQR model, we form the optimality criterion $\hat{F}_Y$ by estimating the involved unknown parameters with the previously described (pilot) data. To identify the optimal sampling schedule for this case, we discretize the domain $\mathcal{T}$ with a regular grid of size 0.5 years. Over the discretized domain, we use the BESS approach with the PSS algorithm to find BESS designs. The

algorithmic parameters of the PSS algorithm are set to their default values described in [21].

Our obtained optimal sampling plans, and their approximated relative efficiencies $\widehat{RE}_Y$ are reported in Table 4. For instance, with $p = 3$, the BESS design suggests to measure the ADAS-cog rating at age 63, 69 and 88 for predicting the HCI score. This design has $\widehat{RE}_Y = 0.7779$. Moreover, as we increase the number of observations p for each patient, the relative efficiency also increases. This result can be expected from the previous discussions, such as those in Section 2. But, we also note here that the increment in the relative efficiency decreases with p. For example, the achieved relative efficiency of the design with $p = 3$ is about 0.15 higher than that of the optimal 2-point design. However, this improvement becomes less than 0.02 when moving from $p = 5$ to $p = 6$. Similar observations are also reported in the literature, and are utilized in selecting p. Some discussions and methods for the selection of p can be found in [17] and [21].

5.2 Berkeley Growth Data

Berkeley growth data consist of 54 girls and 39 boys, and the height of each individual was measured 31 times between age 1 and 18 [24]. We consider the heights from age 1 to 11 as predictor functions to predict the height at age 18. This data set is also used in other studies such as [15], in which optimal designs with a different study objective are obtained. For the purpose of our study, we only use part of the data. The collected data from age 1 to 11 are shown in Figure 4.

For each of the two gender groups, we compare the $\hat{R}^2_M$ in Eq.(9) for the FSLR and FQR models. For the female group, $\hat{R}^2_Q$ and $\hat{R}^2_L$ are 0.8304 and 0.6337, respectively, whereas the male group has $\hat{R}^2_Q = 0.8549$ and $\hat{R}^2_L =$

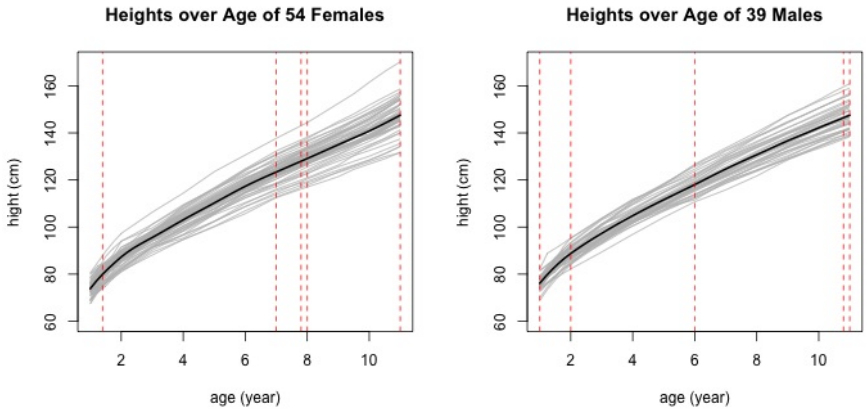

Fig. 4 Berkeley growth data with mean functions. Dashed vertical lines represent the optimal sampling schedules with $p = 5$ observations.

0.8062. Clearly, the FQR models fit better to the data than the FSLR model, especially for the female group. For both groups, the association between the predictor function and the scalar response tends to be strong.

We use the BESS approach to obtain the optimal sampling schedules under the FSLR and the FQR models for $p = 2$ to 6. Table 5 presents the achieved relative efficiencies $\widehat{RE}_Y$, evaluated under the FQR model. The designs obtained under FQR clearly have higher relative efficiencies than those found under FSLR. This is especially true for the female group, which in part, reflects the observations we made previously about the model fitting. There is again an improvement in the relative efficiency when p increases. This improvement is observed for both FQR and FSLR designs, but we see a greater improvement for the FQR designs than the FSLR designs. Partly because of this, the difference in $\widehat{RE}_Y$ between the FQR and FSLR designs tends to increase with p.

Table 5 Relative efficiencies of optimal designs under quadratic and linear models when the number of observations $p = 2, ..., 6$. opt.quad and opt.lin represent the quadratic and linear models, respectively.

	Female		Male	
	FQR	FSLR	FQR	FSLR
$p = 2$	0.8938	0.8632	0.9186	0.8998
$p = 3$	0.9180	0.8666	0.9302	0.9302
$p = 4$	0.9317	0.8832	0.9399	0.9336
$p = 5$	0.9377	0.8846	0.9517	0.9349
$p = 6$	0.9434	0.8861	0.9564	0.9409

6 Discussion

We propose methods for selecting optimal sampling schedule design for functional regression. We derive a new optimality criterion for evaluating competing designs under the FQR models, and provide some important properties of this criterion. Our criterion is reduced to the one for the FSLR model used in the previous studies when the quadratic term is not needed in the model. But for cases where a quadratic relationship between Y and $X(t)$ is present, the designs obtained with our criterion outperform those for FSLR. This observation is consistent across all the scenarios that we had studied, even when the Gaussian assumption is violated.

The derived criterion involves some unknown parameters, such as the eigenfunctions, eigenvalues, and the needed number of these eigenpairs for capturing a sufficiently large portion (e.g., 99%) of the variability of $X(t)$. Previous studies mainly focused on the locally optimal design approach by using the available pilot data to estimate the needed parameters. When the parameter estimates equal their true values, the obtained designs are truly optimal. However, in many realistic situations, the performance of the locally optimal designs may be rather unstable (see also Subsection 4.2). To alleviate this design dependence issue, our second contribution in this paper is the proposed

BESS algorithm for countering the parameter estimation uncertainty in functional regression. As demonstrated in Subsection 4.2, the BESS method can make a better use of the available pilot data to generate designs that yield a higher and more robust performance than the locally optimal design approach. The improvement is significant with small-/medium-sized pilot data sets. The proposed design approach is thus expected to be especially useful for cases where data collection is expensive and/or difficult, as the situation considered in this and previous works. Nevertheless, the performance of the BESS designs also tends to be comparable to, or slightly better than the locally optimal designs when the latter designs are expected to perform well (e.g., with a large pilot data set). We also note that the proposed BESS approach does not require much additional computing time. Obtaining one BESS design in our simulation studies requires no more than one minute on a computer with a 2.8 GHz Quad-Core Intel Core i7 processor. In addition, this approach can also be easily applied to finding high-quality designs for various objectives such as for recovering a predictor function, predicting a scalar/functional response, and a mixture of such objectives.

We also note that the optimality criterion (8) for the BESS approach does not assume a fixed number, K, of eigenpairs. Instead, K is estimated from each bootstrap sample, and is allowed to vary across these samples. This criterion thus also takes the uncertainty of estimating K into account. Nevertheless, the proportion of the variability of $X(t)$ to be captured by the K eigenpairs is fixed. In our simulations, we set this proportion to 99%, but a different value may be considered.

Much research remains to be done in the design of experiments for functional data analysis. A future research of interest is to extend the proposed method to other functional regression models with rather complex structures, such as polynomial regression, multiple regression involving two or more pre-

dictor functions, and function-on-function regression. In addition, a finite K is typically used in (4) for pragmatic reasons, which can lead to a potential bias in predicting Y. Having a design approach to help to reduce this bias should be of interest. Moreover, we assume that the data collected from the pilot study allow the estimation of the underlying model. It is known that the bagging strategy, while useful for improving unstable estimations, does not remedy the problem when the data give serious estimation biases. Selecting a good design to allow the collection of high quality pilot data is thus another important issue to address.

Acknowledgments

This research was supported by National Science Foundation grant CMMI-17-26445 and DMS-13-52213. The authors are grateful to anonymous reviewers for their comments and suggestions, which resulted in an improvement of the presentation of the paper.

Supporting Information

Supplementary materials of additional simulation results can be found online.

Appendix A. Derivation of Optimality Criterion

From Equations (3) and (4),

$$E(Y|X) = \alpha + \sum_{k=1}^{K} \beta_k \zeta_k + \sum_{k=1}^{K}\sum_{l=1}^{K} \gamma_{kl}\zeta_k\zeta_l = \alpha + \boldsymbol{b}^T\boldsymbol{\zeta} + \boldsymbol{\zeta}^T \mathrm{G}\boldsymbol{\zeta}, \text{ and}$$
$$E(E(Y|X)|\boldsymbol{U}) = \alpha + \sum_{k=1}^{K} \beta_k \tilde{\zeta}_k + \sum_{k=1}^{K}\sum_{l=1}^{K} \gamma_{kl}\widetilde{\zeta_k\zeta_l} = \alpha + \boldsymbol{b}^T\tilde{\boldsymbol{\zeta}} + \mathrm{tr}(\mathrm{G}\widetilde{\boldsymbol{\zeta}\boldsymbol{\zeta}^T}),$$

where $\boldsymbol{\zeta} = (\zeta_1, ..., \zeta_K)^T$, $\tilde{\boldsymbol{\zeta}} = (\tilde{\zeta}_1, ..., \tilde{\zeta}_K)^T$ and $\widetilde{\boldsymbol{\zeta}\boldsymbol{\zeta}^T} = \tilde{\boldsymbol{\zeta}}\tilde{\boldsymbol{\zeta}}^T - \mathrm{Cov}(\boldsymbol{\zeta}) + \mathrm{Cov}(\tilde{\boldsymbol{\zeta}})$; see also [29].

$$E\Big(E(Y|X) - E(E(Y|X)|\boldsymbol{U})\Big)^2 = \mathrm{Var}\Big(E(Y|X) - E(E(Y|X)|\boldsymbol{U})\Big)$$
$$= \mathrm{Var}\Big(E(Y|X)\Big) + \mathrm{Var}\Big(E(E(Y|X)|\boldsymbol{U})\Big) - 2\mathrm{Cov}\Big(E(Y|X), E(E(Y|X)|\boldsymbol{U})\Big)$$

Each term is obtained below. We note that the third central moment of a normal distribution is 0 and $\boldsymbol{\zeta}$ follows a multivariate normal distribution with mean 0 and variance-covariance matrix R. We then have:

$$\mathrm{Var}\Big(E(Y|X)\Big) = \mathrm{Var}(\alpha + \boldsymbol{b}^T\boldsymbol{\zeta} + \boldsymbol{\zeta}^T\mathrm{G}\boldsymbol{\zeta})$$
$$= \mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T R + 2(\mathrm{G}R)^2\big].$$

We note that $\tilde{\boldsymbol{\zeta}} = R\Psi^T\Gamma_*^{-1}(\boldsymbol{U} - \boldsymbol{\mu}_X) = H\Gamma_*^{-1}(\boldsymbol{U} - \boldsymbol{\mu}_X)$ follows a multivariate normal distribution with mean 0 and variance-covariance matrix $H\Gamma_*^{-1}H^T$.

$$\mathrm{Var}\Big(E(E(Y|X)|\boldsymbol{U})\Big) = \mathrm{Var}\Big(\alpha + \boldsymbol{b}^T\tilde{\boldsymbol{\zeta}} + \mathrm{tr}\big[\mathrm{G}\widetilde{\boldsymbol{\zeta}\boldsymbol{\zeta}^T}\big]\Big)$$
$$= \mathrm{Var}\Big(\boldsymbol{b}^T\tilde{\boldsymbol{\zeta}} + \mathrm{tr}\big[\mathrm{G}(\tilde{\boldsymbol{\zeta}}\tilde{\boldsymbol{\zeta}}^T + R - H\Gamma_*^{-1}H^T)\big]\Big)$$
$$= \mathrm{Var}(\boldsymbol{b}^T\tilde{\boldsymbol{\zeta}}) + \mathrm{Var}(\tilde{\boldsymbol{\zeta}}^T\mathrm{G}\tilde{\boldsymbol{\zeta}}) + 2\mathrm{Cov}(\boldsymbol{b}^T\tilde{\boldsymbol{\zeta}}, \tilde{\boldsymbol{\zeta}}^T\mathrm{G}\tilde{\boldsymbol{\zeta}})$$
$$= \mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T H\Gamma_*^{-1}H^T + 2(\mathrm{G}H\Gamma_*^{-1}H^T)^2\big]$$

We also note that $\boldsymbol{U} = \boldsymbol{\mu}_X + \Psi\boldsymbol{\zeta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{\zeta}$ and follows a multivariate normal distribution with mean 0 and variance-covariance matrix $\sigma_X^2 I_p$. According to [1],

$$\mathrm{Cov}\Big(E(Y|X), E(E(Y|X)|\boldsymbol{U})\Big)$$
$$= \mathrm{Cov}\Big(\boldsymbol{b}^T\boldsymbol{\zeta} + \boldsymbol{\zeta}^T\mathrm{G}\boldsymbol{\zeta}, \boldsymbol{b}^T\tilde{\boldsymbol{\zeta}} + \tilde{\boldsymbol{\zeta}}^T\mathrm{G}\tilde{\boldsymbol{\zeta}}\Big)$$
$$= \mathrm{Cov}\Big(\boldsymbol{b}^T\boldsymbol{\zeta}, \boldsymbol{b}^T H\Gamma_*^{-1}\Psi\boldsymbol{\zeta}\Big) + \mathrm{Cov}\Big(\boldsymbol{\zeta}^T\mathrm{G}\boldsymbol{\zeta}, \boldsymbol{\zeta}^T\Psi^T\Gamma_*^{-1}H^T\mathrm{G}H\Gamma_*^{-1}\Psi\boldsymbol{\zeta}\Big)$$
$$= \mathrm{tr}\big[\boldsymbol{b}\boldsymbol{b}^T H\Gamma_*^{-1}H^T + 2(\mathrm{G}H\Gamma_*^{-1}H^T)^2\big]$$

Appendix B. Proofs of Theorems

We denote H and Γ_* are observed at time points $\mathbf{t}$, and $\tilde{H}$ and $\tilde{\Gamma}_*$ are observed at time points $\tilde{\mathbf{t}}$. Park *et al.* [17] have proved theorems under the regularity conditions in [4] when the functional linear regression is considered. Moreover, without loss of generality, we assume that $\mathcal{T} = [0, 1]$.

Proof of Theorem 1

For any positive semidefinite matrix B, $\mathrm{tr}\big[BH\Gamma_*^{-1}H^T\big] \le \mathrm{tr}\big[B\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T\big]$, and $\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T - H\Gamma_*^{-1}H^T$ is a positive semidefinite matrix [17]. Therefore, it suffices to show that $\mathrm{tr}\big[(\mathrm{G}H\Gamma_*^{-1}H^T)^2\big] \le \mathrm{tr}\big[(\mathrm{G}\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T)^2\big]$. Since $\mathrm{G}(\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T + H\Gamma_*^{-1}H^T)\mathrm{G}$ and $\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T - H\Gamma_*^{-1}H^T$ are positive semidefinite matrices, there exist matrices P and Q such that $\mathrm{G}(\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T + H\Gamma_*^{-1}H^T)\mathrm{G} = PP^T$ and $\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T - H\Gamma_*^{-1}H^T = QQ^T$.

$$\begin{aligned}
&\mathrm{tr}\big[(\mathrm{G}\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T)^2\big] - \mathrm{tr}\big[(\mathrm{G}H\Gamma_*^{-1}H^T)^2\big] = \mathrm{tr}\big[(\mathrm{G}\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T)^2 - (\mathrm{G}H\Gamma_*^{-1}H^T)^2\big] \\
&\quad = \mathrm{tr}\big[(\mathrm{G}\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T + \mathrm{G}H\Gamma_*^{-1}H^T)(\mathrm{G}\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T - \mathrm{G}H\Gamma_*^{-1}H^T)\big] \\
&\quad = \mathrm{tr}\big[\mathrm{G}(\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T + H\Gamma_*^{-1}H^T)\mathrm{G}(\tilde{H}\tilde{\Gamma}_*^{-1}\tilde{H}^T - H\Gamma_*^{-1}H^T)\big] \\
&\quad = \mathrm{tr}\big(PP^TQQ^T\big) = ||P^TQ||_F^2 \ge 0.
\end{aligned}$$

Lemma 1 *Let $A, B \in \mathbb{R}^{N\times N}$ be symmetric and positive semidefinite matrices for any positive integer N. Then,* $\mathrm{tr}(AB) \le \mathrm{tr}(A)\mathrm{tr}(B)$.

Proof of Lemma 1

Since A and B are positive semidefinite matrices, there exist matrices P and Q such that $A = PP^T$ and $B = QQ^T$. Then,

$$\begin{aligned}
\mathrm{tr}(AB) &= \mathrm{tr}(PP^TQQ^T) = \mathrm{tr}(P^TQ(P^TQ)^T) = ||P^TQ||_F^2 \\
&\le ||P^T||_F^2||Q||_F^2 = \mathrm{tr}(PP^T)\mathrm{tr}(QQ^T) = \mathrm{tr}(A)\mathrm{tr}(B).
\end{aligned}$$

Proof of Theorem 2

Park *et al.* [17] have shown that $\lim_{p\to\infty}\mathrm{tr}(BH\Gamma_*^{-1}H^T) = \mathrm{tr}(BR)$, for any positive semidefinite matrix B. Thus, it suffices to show that $\lim_{p\to\infty}\mathrm{tr}\big[(\mathrm{G}H\Gamma_*^{-1}H^T)^2\big] = \mathrm{tr}\big[(\mathrm{G}R)^2\big]$. Since $\lim_{p\to\infty}\mathrm{tr}(H\Gamma_*^{-1}H^T) = \mathrm{tr}(R)$, for any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $\mathrm{tr}(R -$

$H\Gamma_*^{-1}H^T) < \epsilon/(\mathrm{tr}(\mathrm{G}^2)2\rho_1 K)$ when $p \geq N$. For given $\varepsilon > 0$, if $p \geq N$, then by Lemma 1,

$$\begin{aligned}
\mathrm{tr}[(GR)^2] - \mathrm{tr}[(GH\Gamma_*^{-1}H^T)^2] &= \mathrm{tr}\big[(\mathrm{G}R)^2 - (\mathrm{G}H\Gamma_*^{-1}H^T)^2\big] \\
&= \mathrm{tr}\big[(\mathrm{G}R + \mathrm{G}H\Gamma_*^{-1}H^T)(\mathrm{G}R - \mathrm{G}H\Gamma_*^{-1}H^T)\big] \\
&= \mathrm{tr}\big[\mathrm{G}(R + H\Gamma_*^{-1}H^T)\mathrm{G}(R - H\Gamma_*^{-1}H^T)\big] \\
&\leq \mathrm{tr}\big[\mathrm{G}(R + H\Gamma_*^{-1}H^T)\mathrm{G}\big]\ \mathrm{tr}(R - H\Gamma_*^{-1}H^T) \\
&\leq \mathrm{tr}(\mathrm{G}^2)\ 2\mathrm{tr}(R)\ \mathrm{tr}(R - H\Gamma_*^{-1}H^T) \\
&\leq \mathrm{tr}(\mathrm{G}^2)\ 2\rho_1 K\ \mathrm{tr}(R - H\Gamma_*^{-1}H^T) < \epsilon.
\end{aligned}$$

Appendix C. Pseudo Code for PSS Algorithm

Algorithm: Probabilistic subset search algorithm

INPUT: Candidate points $\mathcal{X}$; initial design d_p^0; subset size l
OUTPUT: Optimal sampling schedule d_p

Set $k \leftarrow 1$
Calculate weights π_{s_i}
while d_p^{k-1} does not satisfy the stopping rule **do**
Choose a subset S_k whose elements are randomly selected without replacement from $\mathcal{X} - d_p^{k-1}$ with probability proportional to π_{s_i}
$\mathcal{X}_k = S_k \cup d_p^{k-1}$ and $\Xi_k = \{d_p : d_p(j) \in \mathcal{X}_k, j = 1, ..., p\}$
Compute $F_Y(\cdot)$ for the $\binom{p+l}{p}$ designs in Ξ_k
Find $d_p^* = \arg\max_{d_p \in \Xi_k} F_Y(d_p)$
$d_p^k \leftarrow d_p^*$
$k \leftarrow k + 1$
end while

Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Bao, Y. and Ullah, A. (2010). Expectation of quadratic forms in normal and nonnormal variables with applications. *Journal of Statistical Planning and Inference* **140,** 1193-1205.
2. Breiman, L. (1996). Bagging predictors. *Machine Learning* **24,** 123-140.
3. Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The Annals of Statistics* **30,** 927-961.
4. Bunea, F. and Xiao, L. (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *Bernoulli* **21,** 1200-1230.
5. Chen, K., Ayutyanont, N., Langbaum, J. B., Fleisher, A. S., Reschke, C., Lee, W., et al. (2011). Characterizing Alzheimer's disease using a hypometabolic convergence index. *NeuroImage* **56,** 52-60.
6. Chen, Y., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K., Ji, H., et al. (2019). *fdapace: Functional data analysis and empirical dynamics* R package version 0.5.1.

7. Chernoff, H. (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics* **24,** 586-602.

8. Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators.* UK: John Wiley & Sons Inc.

9. James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87,** 587-602.

10. Ji, H. and Müller, H. G. (2017). Optimal designs for longitudinal and functional data. *Journal of the Royal Statistical Society, Series B* **79,** 859-876.

11. Jiang, C. R., Aston, J. A., and Wang, J. L. (2009). Smoothing dynamic positron emission tomography time courses using functional principal components. *NeuroImage* **47,** 184-193.

12. Kolibas, E., Korinkova, V., Novotny, V., Vajdickova, K., and Hunakova, D. (2000). ADAS-cog (Alzheimer's disease assessment scale-cognitive subscale) validation of the Slovak version. *European Psychiatry* **15,** 443.

13. Lai, T. L., Shih, M. C., and Wong, S. P. (2006). A new approach to modeling covariate effects and individualization in population pharmacokinetics-pharmacodynamics. *Journal of Pharmacokinetics and Pharmacodynamics* **33,** 49-74.

14. Li, C. (2017). *Statistical Methods for Functional and Complex Data.* North Carolina State University. Ph.D. Dissertation.

15. Li, C. and L. Xiao (2020). *Statistical Methods for Functional and Complex Data.* Optimal design for classification of functional data. *Canadian Journal of Statistics* **48,** 285-307.

16. Müller H. G., Sen, R., and Stadtmüller, U. (2011). Functional data analysis for volatility. *Journal of Econometrics* **165,** 233-245.

17. Park, S. Y., Xiao, L., Willbur, J. D., Staicu, A. M., and Jumbe, N. L. (2018). A joint design for functional data with application to scheduling ultrasound scans. *Computational Statistics & Data Analysis* **122,** 101-114.

18. Peng, J. and Paul, D. (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphic Statistics* **18,** 995-1015.

19. Ramsay, J. O. and Ramsay J. B. (2002). Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of Econometrics* **107,** 327-344.

20. Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis.* New York: Springer.

21. Rha, H., Kao, M. H., and Pan, R. (2020). Design optimal sampling plans for functional regression models. *Computational Statistics & Data Analysis* **146,** 106925.

22. Saleh, M., Kao, M. H., and Pan, R. (2017). Design D-optimal event-related functional magnetic resonance imaging experiments. *Journal of the Royal Statistical Society, Series C* **66,** 73-91.

23. Sørensen, H., and Goldsmith, J., and Sangalli, L. M. (2013). An introduction with medical applications to functional data analysis. *Statistics in Medicine* **32,** 5222-5240.

24. Tuddenham, R. D. and Snyder, M. M. (1954). Physical growth of California boys and girls from birth to age 18. *University of California Publications in Child Development* **1,** 183-364.

25. West, R. M., Harris, K., Giilthorpe, M. S., Tolman, C., and Will, E. J. (2007). Functional data analysis applied to a randomized controlled clinical trial in hemodialysis patients describes the variability of patient responses in the control of renal anemia. *Journal of the American Society of Nephrology* **18,** 2371-2376.

26. Wu, M., Diez-Roux, A., Raghunathan, T. E., and Sánchez, B. N. (2018). FPCA-based method to select optimal sampling schedules that capture between-subject variability in longitudinal studies. *Biometrics* **74,** 229-238.

27. Xiao, L., Li, C., Checkley, W., and Crainiceanu, C. (2018). Fast covariance estimation for sparse functional data. *Statistics and Computing* **28,** 511-522.

28. Yao, F., Liu, B., Tao, W., Wu, S., Yang, N. T., Yang, W. et al. (2015). *PACE: Principal analysis by conditional expectation* MATLAB package version 2.17.

29. Yao, F. and Müller, H. G. (2010). Functional quadratic regression. *Biometrika* **97,** 49-64.

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

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