ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 3, *pp.* 859–876

Optimal designs for longitudinal and functional data

Hao Ji and Hans-Georg Müller

University of California at Davis, USA

[Received February 2015. Final revision May 2016]

Summary. We propose novel optimal designs for longitudinal data for the common situation where the resources for longitudinal data collection are limited, by determining the optimal locations in time where measurements should be taken. As for all optimal designs, some prior information is needed to implement the optimal designs proposed. We demonstrate that this prior information may come from a pilot longitudinal study that has irregularly measured and noisy measurements, where for each subject one has available a small random number of repeated measurements that are randomly located on the domain. A second possibility of interest is that a pilot study consists of densely measured functional data and one intends to take only a few measurements at strategically placed locations in the domain for the future collection of similar data. We construct optimal designs by targeting two criteria: optimal designs to recover the unknown underlying smooth random trajectory for each subject from a few optimally placed measurements such that squared prediction errors are minimized; optimal designs that minimize prediction errors for functional linear regression with functional or longitudinal predictors and scalar responses, again from a few optimally placed measurements. The optimal designs proposed address the need for sparse data collection when planning longitudinal studies, by taking advantage of the close connections between longitudinal and functional data analysis. We demonstrate in simulations that the designs perform considerably better than randomly chosen design points and include a motivating data example from the Baltimore Longitudinal Study of Aging. The designs are shown to have an asymptotic optimality property.

Keywords: Asymptotics; Coefficient of determination; Functional data analysis; Functional principal components; Gaussian process; Karhunen–Loève expansion; Longitudinal data; Prediction error; Sparse design

1. Introduction

Functional data analysis has become increasingly useful in various fields. In many applications, especially in longitudinal studies, often only a few repeated measurements can be obtained for each subject or item, owing to cost or logistical constraints that limit the number of measurements. In some functional or longitudinal data where the recordings are sparse and have been taken at irregular time points, functional data analysis methodology has proved useful to infer covariance structure and trajectories (Staniswalis and Lee, 1998; Yao *et al.*, 2005a; Li and Hsing, 2010). Although, ideally, longitudinal and functional data would be measured on a dense grid, in practical studies we usually encounter constraints on data collection. It is then of interest to have criteria and principles to determine where on the domain (usually but not necessarily a time interval) we should place a given number of measurements to minimize prediction errors when recovering the unobserved trajectories for each subject or to predict a response that is associated with each longitudinal trajectory. Sparse sampling is also of interest in applications where we have available densely sampled functional data but can sample at only a few important

Address for correspondence: Hao Ji, Department of Statistics, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA.
E-mail: haoji@ucdavis.edu

1369–7412/17/79859

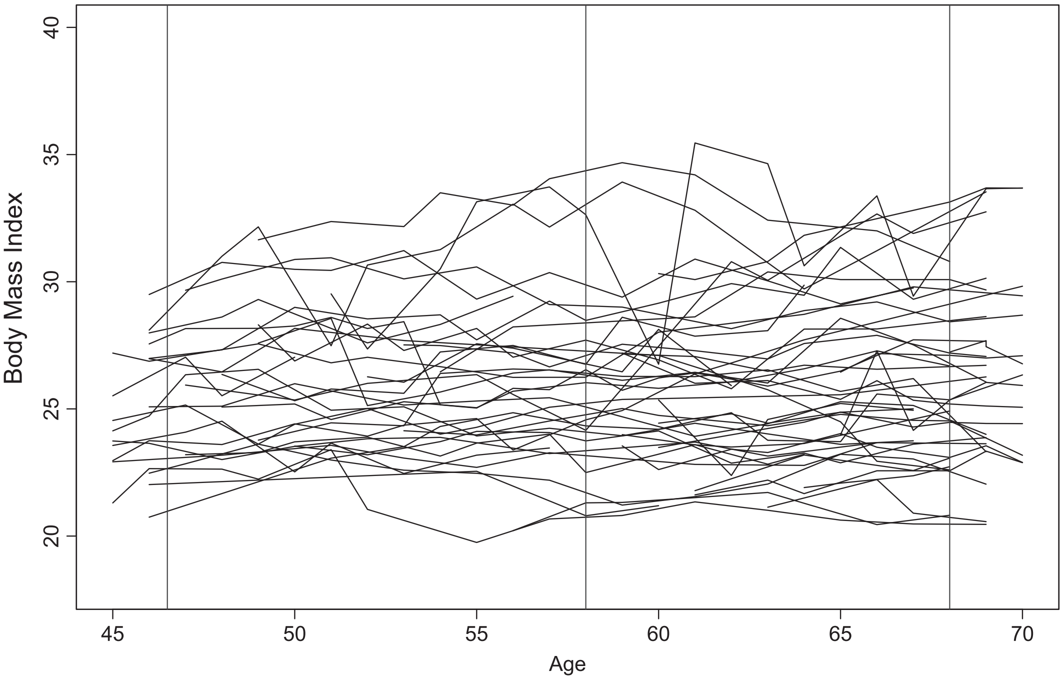


Fig. 1. Spaghetti plot for a subset of BMI profiles observed in the BLSA: |, estimated locations of the optimal design points for recovering BMI trajectories, among designs with three measurement locations

locations in future longitudinal data collection. The question that we address here is where these locations should be. Answering these questions is for example of keen interest to determine optimal monitoring schedules for children in developing countries, aiming to recover their growth trajectories from sparse measurements to assess growth stunting and faltering.

Several previous studies have discussed methods and algorithms for finding optimal designs in the dense functional data case where data are sampled on a dense regular grid (Ferraty *et al.*, 2010; Delaigle *et al.*, 2012) or belong to a special family of functions (McKeague and Sen, 2010), where the emphasis has been on non-parametric functional regression and classification. However, these methods cannot be extended to the case of sparse functional or longitudinal data, which is a case of crucial interest, as the design selection impacts the planning of longitudinal studies that are often cost intensive. A crucial design feature is where to place a limited number of future longitudinal observations. Whereas the connections between longitudinal and non-parametric approaches are being increasingly studied (Guo, 2004; Xiang *et al.*, 2013) and there are also studies on designs for classical parametric longitudinal models (Mentre *et al.*, 1997; Anisimov *et al.*, 2007) and for random processes and fields (Zagoraiou and Baldi Antognini, 2009; Fedorov and Leonov, 2013), to our knowledge, there is no previous work on optimal designs for longitudinal studies when the underlying longitudinal trajectories are viewed as smooth random functions. The optimal designs proposed fill this gap and are found to perform well in simulations (see Section 5).

Our approach is motivated by the idea of design selection for longitudinal studies such as the well-known Baltimore Longitudinal Study of Aging (BLSA) (Shock *et al.*, 1984; Pearson *et al.*, 1997), where (among other variables) body mass index (BMI) profiles are measured sparsely in time. Currently available data essentially feature random timing of the measurements (Fig. 1).

Our methods will be useful to determine optimal designs for a follow-up study or to recruit new subjects, where we develop optimal designs for

(a) recovering the unknown smooth underlying trajectories, as these cannot be easily obtained with non-parametric methods because of the sparseness of the measurements, and
(b) for predicting scalar responses in functional linear models.

As we shall demonstrate, sparsely and irregularly sampled data from a pilot study suffice to construct consistent estimates for the optimal designs. A design resulting from our methodology is depicted in Fig. 1, where the three highlighted locations are optimally placed to recover the underlying smooth BMI trajectories when one is constrained to select just three measurements of BMI in future studies.

The vector of function values $\mathbf{X}=(X(t_1),\ldots,X(t_p))^{\mathrm{T}}$ of a generic process X at design points $\mathbf{t}=(t_1,\ldots,t_p)^{\mathrm{T}}$, where T denotes transpose, is not observed in practice, as observations are contaminated with measurement errors. We refer to the observed values at the design points as $\mathbf{U}=(U(t_1),\ldots,U(t_p))^{\mathrm{T}}$, where

$$\left.\begin{aligned} U(t_j)&=X(t_j)+e_j,\\ E(e_j)&=0,\\ \operatorname{var}(e_j)&=\sigma^2, \qquad j=1,\ldots,p, \end{aligned}\right\} \tag{1}$$

and errors e_j are independent. For trajectory recovery, we use best linear predictors that we denote by B (Rice and Wu, 2001):

$$B\{X(t)|\mathbf{U}\}=\mu(t)+\operatorname{cov}\{X(t),\mathbf{U}\}\operatorname{cov}(\mathbf{U})^{-1}(\mathbf{U}-\boldsymbol{\mu}), \qquad \mu(t)=E\{X(t)\}, \tag{2}$$

with $\boldsymbol{\mu}=E(\mathbf{U})=(\mu(t_1),\mu(t_2),\ldots,\mu(t_p))^{\mathrm{T}}$; for details of the derivation see the on-line supplement section A.1. Under Gaussian assumptions these are also the best predictors, as then $E\{X(t)|\mathbf{U}\}=B\{X(t)|\mathbf{U}\}$. Optimal designs for trajectory recovery are derived by minimizing the expected squared distance between these best linear predictors and the true trajectories, which turns out to be equivalent to maximizing a generalized coefficient of determination with respect to the design points (for details see the on-line supplement section A.3).

For response prediction, where a functional predictor is coupled with a scalar response, the functional linear model is a classical approach (Cardot *et al.*, 1999; Ramsay and Silverman, 2005),

$$E(Y|X)=\mu_Y+\int_{\mathcal{T}}\beta(t)\,X^{\mathrm{c}}(t)\,\mathrm{d}t, \tag{3}$$

where $\mu_Y=E(Y)$, $\beta(t)$ is the regression coefficient function and $X^{\mathrm{c}}(t)$ is the centred predictor process, i.e. $X^{\mathrm{c}}(t)=X(t)-\mu(t)$. The difficulty in longitudinal designs is that the integral on the right-hand side of model (3) cannot be evaluated, whenever there is sparse sampling of trajectories X_i. Therefore, for sparse designs, we condition on $\mathbf{U}$ on both sides of model (3), giving the sparse version

$$B\{E(Y|X)|\mathbf{U}\}=\mu_Y+\int_{\mathcal{T}}\beta(t)\,B\{X^{\mathrm{c}}(t)|\mathbf{U}\}\,\mathrm{d}t, \tag{4}$$

where again, under the Gaussian assumption, $E\{X^{\mathrm{c}}(t)|\mathbf{U}\}=B\{X^{\mathrm{c}}(t)|\mathbf{U}\}$ and $E\{E(Y|X)|\mathbf{U}\}=B\{E(Y|X)|\mathbf{U}\}$. Then the optimal designs $\mathbf{t}=\{t_1,\ldots,t_p\}$ are defined as the minimizers of the squared prediction error $E[Y-B\{E(Y|X)|\mathbf{U}\}]^2$.

The paper is organized as follows. We discuss optimal designs for trajectory recovery and response prediction in Sections 2 and 3 respectively, followed by estimated optimal designs in

Section 4 and numerical implementations in Section 5, with simulation studies in Section 6. Data analysis examples from various areas are in Section 7, and asymptotic results in Section 8.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Optimal designs for trajectory recovery

For fixed p, consider a generic set of non-random design points $\mathbf{t}=(t_1,\ldots,t_p)^{\mathrm{T}}$, and corresponding values of the underlying process, $\mathbf{X}=(X(t_1),\ldots,X(t_p))^{\mathrm{T}}$, with noisy observations $\mathbf{U}=(U(t_1),\ldots,U(t_p))^{\mathrm{T}}$ and mean vector $\boldsymbol{\mu}=(\mu(t_1),\ldots,\mu(t_p))^{\mathrm{T}}$. We aim to find optimal designs $\mathbf{t}^*_{\mathbf{X}}$ with respect to a target criterion. For the underlying process X we write $\mu(t)=E\{X(t)\}$ for the mean and $\Gamma(s,t)=\mathrm{cov}\{X(s),X(t)\}$ for the autocovariance function, and the covariance matrices $\boldsymbol{\Gamma}=\mathrm{cov}(\mathbf{X})$ and $\boldsymbol{\Gamma}_*=\mathrm{cov}(\mathbf{U})$, where we note that expression (1) implies that $\boldsymbol{\Gamma}_*=\boldsymbol{\Gamma}+\sigma^2 I_p$, with I_p denoting the $p\times p$ identity matrix. With $\mu(t)=E\{X(t)\}$, the best linear predictor in equation (2) becomes

$$\begin{aligned} B\{X(t)|\mathbf{U}\}&=q(t)+\boldsymbol{\alpha}(t)^{\mathrm{T}}\mathbf{U}, \qquad & q(t)&=\mu(t)-\gamma(t)^{\mathrm{T}}\boldsymbol{\Gamma}_*^{-1}\boldsymbol{\mu},\\ \boldsymbol{\alpha}(t)&=\boldsymbol{\Gamma}_*^{-1}\gamma(t), \qquad & \gamma(t)&=\mathrm{cov}\{\mathbf{U},X(t)\}=(\Gamma(t_1,t),\ldots,\Gamma(t_p,t))^{\mathrm{T}}, \end{aligned} \tag{5}$$

where $\gamma(t)$ is a p-dimensional vector of covariances associated with the non-random time points $\mathbf{t}=(t_1,\ldots,t_p)$. Here and in what follows, expectations and covariances are considered to be conditional on designs $\mathbf{t}$.

Our goal is to minimize the mean integrated squared error MISE of recovered trajectories that are obtained with the best linear predictors $B\{X(t)|\mathbf{U}\}$ as a function of the design points $\mathbf{t}=(t_1,\ldots,t_p)^{\mathrm{T}}$ that exert their influence through $\mathbf{X}=(X(t_1),\ldots,X(t_p))^{\mathrm{T}}$, i.e. to minimize

$$\mathrm{MISE}(\mathbf{t})=E\left(\int_{\mathcal{T}}[X(t)-B\{X(t)|\mathbf{U}\}]^2\,\mathrm{d}t\right). \tag{6}$$

The performance of recovered trajectories at a fixed point $t\in\mathcal{T}$ can be quantified by a pointwise coefficient of determination, defined as

$$R^2(t)=\frac{\mathrm{var}[B\{X(t)\}|\mathbf{U}]}{\mathrm{var}\{X(t)\}}. \tag{7}$$

This motivates us to maximize an overall coefficient of determination with respect to the design points. It is easy to see that

$$R^2_X=\frac{\int_{\mathcal{T}}\gamma(t)^{\mathrm{T}}\boldsymbol{\Gamma}_*^{-1}\gamma(t)\,\mathrm{d}t}{\int_{\mathcal{T}}\mathrm{var}\{X(t)\}\,\mathrm{d}t}, \tag{8}$$

is a well-defined coefficient of determination for all $\mathbf{t}\in\mathcal{T}(\delta_0)$ for some $\delta_0>0$, where $\mathcal{T}(\delta_0)$ is defined as

$$\mathcal{T}(\delta_0)=\{(t_1,\ldots,t_p)^{\mathrm{T}}\in\mathcal{T}^p,\mathrm{cov}\{U(t_1),\ldots,U(t_p)\}-\delta_0 I_p \text{ is positive definite}\}. \tag{9}$$

Requiring $\mathbf{t}\in\mathcal{T}(\delta_0)$ ensures uniform matrix invertibility across designs. We assume that the true optimal design also satisfies $\mathbf{t}^0_X\in\mathcal{T}(\delta_0)$. In the on-line supplement section A.3 we show that minimizing MISE in dependence on $\mathbf{t}$ as in equation (6) is equivalent to maximizing R^2_X in equation (8), which is in turn equivalent to finding

$$\mathbf{t}_X^* = (t_{X_1}^*, \ldots, t_{X_p}^*)^{\mathrm{T}} = \underset{(t_1,\ldots,t_p)^{\mathrm{T}} \in \mathcal{T}(\delta_0)}{\arg\max} \int_{\mathcal{T}} \gamma(t)^{\mathrm{T}} \boldsymbol{\Gamma}_*^{-1} \gamma(t)\,\mathrm{d}t. \tag{10}$$

3. Optimal designs for predicting scalar responses

We assume here the same setting and conditions as in the previous section and aim at finding optimal designs $\mathbf{t}_Y^*$ with respect to a target criterion that is specific for the functional linear model (3).

The best linear predictor in equation (4), using expression (5), is seen to be

$$\begin{aligned} B\{E(Y|X)|\mathbf{U}\} &= \mu_Y + \int_{\mathcal{T}} \beta(t)\, B\{X^{\mathrm{c}}(t)|\mathbf{U}\}\,\mathrm{d}t \\ &= \mu_Y + \int_{\mathcal{T}} \beta(t)\, \boldsymbol{\alpha}(t)^{\mathrm{T}} (\mathbf{U} - \boldsymbol{\mu})\,\mathrm{d}t = \mu_Y + \boldsymbol{\beta}_p^{\mathrm{T}} (\mathbf{U} - \boldsymbol{\mu}), \end{aligned} \tag{11}$$

with $\boldsymbol{\beta}_p = \int_{\mathcal{T}} \beta(t)\, \boldsymbol{\alpha}(t)\,\mathrm{d}t$. A basic tool for the following derivations is the Karhunen–Loève expansion of square integrable random processes:

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} \zeta_k\, \psi_k(t), \tag{12}$$

with eigenfunctions $\psi_k, k = 1, 2, \ldots$, of the covariance operator of X, and uncorrelated (independent in the Gaussian case) functional principal components $\zeta_k, k = 1, 2, \ldots$. The eigenfunctions form an orthogonal basis of the space generated by X and we have the covariance expansion

$$\Gamma(s, t) = \mathrm{cov}\{X(s), X(t)\} = \sum_{k=1}^{\infty} \rho_k\, \psi_k(s)\, \psi_k(t), \tag{13}$$

where the eigenvalues ρ_k of the covariance operator are positive and ordered, $\rho_1 > \rho_2 > \ldots$. The function principal components satisfy $E(\zeta_k) = 0$ and $\mathrm{var}(\zeta_k) = \rho_k$ for all k.

The regression coefficient function $\beta(t)$ of the functional linear model (3) can be expanded in the eigenbasis representation (He *et al.*, 2000), with convergence under mild regularity conditions:

$$\beta(t) = \sum_{k=1}^{\infty} \frac{E(\zeta_k Y)}{E(\zeta_k^2)}\, \psi_k(t). \tag{14}$$

Observing that for the cross-covariance function

$$C(t) = \mathrm{cov}\{X(t), Y\} = \mathrm{cov}\left\{ \sum_{k=1}^{\infty} \zeta_k\, \psi_k(t), Y \right\} = \sum_{k=1}^{\infty} E(\zeta_k Y)\, \psi_k(t) \tag{15}$$

and, using the orthonormality of the eigenfunctions ψ_k, it is easy to see that $\sigma_k = E(\zeta_k Y)$ can be written as

$$\sigma_k = \int_{\mathcal{T}} C(t)\, \psi_k(t)\,\mathrm{d}t. \tag{16}$$

By equations (5), (14) and (16), we have, with $\boldsymbol{\psi}_k = (\psi_k(t_1), \ldots, \psi_k(t_p))^{\mathrm{T}}$ and $\mathbf{C} = (C(t_1), \ldots, C(t_p))^{\mathrm{T}}$,

$$\begin{aligned} \boldsymbol{\beta}_p &= \int_{\mathcal{T}} \beta(t)\, \boldsymbol{\alpha}(t)\,\mathrm{d}t = \boldsymbol{\Gamma}_*^{-1} \int_{\mathcal{T}} \left\{ \sum_{k=1}^{\infty} \frac{\sigma_k}{\rho_k}\, \psi_k(t) \right\} \left\{ \sum_{k=1}^{\infty} \rho_k\, \psi_k(t)\, \boldsymbol{\psi}_k \right\} \mathrm{d}t \\ &= \boldsymbol{\Gamma}_*^{-1} \sum_{k=1}^{\infty} \sigma_k \boldsymbol{\psi}_k = \boldsymbol{\Gamma}_*^{-1} \mathbf{C}. \end{aligned} \tag{17}$$

Similar to trajectory recovery in the previous section, we propose to minimize the prediction error

$$E[Y - B\{E(Y|X)|\mathbf{U}\}]^2 = E[Y - \mu_Y - \boldsymbol{\beta}_p^{\mathrm{T}}(\mathbf{U} - \boldsymbol{\mu}_{\mathbf{X}})]^2. \tag{18}$$

This is shown in the on-line supplement section A.3 to be equivalent to maximizing the following coefficient of determination R_Y^2 that quantifies prediction power:

$$R_Y^2 = \frac{\mathrm{var}[B\{E(Y|X)|\mathbf{U}\}]}{\mathrm{var}(Y)}, \tag{19}$$

where we assume that the true optimal designs for regression case $\mathbf{t}_Y^0$ lie in $\mathcal{T}(\delta_0)$.

To maximize R_Y^2, it is equivalent to find

$$\mathbf{t}_Y^* = \underset{(t_1,\ldots,t_p)^{\mathrm{T}} \in \mathcal{T}(\delta_0)}{\arg\max}\ \mathrm{var}[B\{E(Y|X)|\mathbf{U}\}] = \underset{(t_1,\ldots,t_p)^{\mathrm{T}} \in \mathcal{T}(\delta_0)}{\arg\max}\ \boldsymbol{\beta}_p^{\mathrm{T}}\boldsymbol{\Gamma}_*\boldsymbol{\beta}_p.$$

Therefore, by equation (17), the optimization criterion can be simplified to

$$\mathbf{t}_Y^* = \underset{(t_1,\ldots,t_p)^{\mathrm{T}} \in \mathcal{T}(\delta_0)}{\arg\max}\ \mathbf{C}^{\mathrm{T}}\boldsymbol{\Gamma}_*^{-1}\mathbf{C}. \tag{20}$$

4. Estimated optimal designs

Although the population optimal designs were derived in the previous sections, in practice they must be estimated from available data. The available observations are

$$U_{ij} = X_i(t_{ij}) + e_{ij}, \qquad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant m_i, \tag{21}$$

$$Y_i = \mu_Y + \int_{\mathcal{T}} \beta(t)\, X_i^{\mathrm{c}}(t)\, \mathrm{d}t + \epsilon_i, \qquad 1 \leqslant i \leqslant n, \tag{22}$$

where expression (22) applies only to the prediction scenario. Here (X_i, Y_i), $i = 1, \ldots, n$, are independent realizations of (X, Y), with m_i the number of observed function values for each subject, the t_{ij} are randomly located time points on $\mathcal{T}$ with density function $f_T(\cdot)$ and the e_{ij} and the ϵ_{ij} are random errors with zero mean and variance σ^2 and σ_Y^2 respectively. We assume throughout that the (X_i, Y_i), the e_{ij} and the ϵ_i are all independent. A notable feature of this data model is that it includes noise not only in the responses Y_i but also in the recordings of the random trajectories.

From the data, estimates of mean function $\mu(t)$, autocovariance function $\Gamma(s,t)$ and cross-covariance function $C(t)$ are obtained on a user-defined fine grid covering $\mathcal{T}$ and these are denoted as $\hat{\mu}(t)$, $\hat{\Gamma}(s,t)$ and $\hat{C}(t)$ respectively. Consistent estimates of these quantities from the pilot study are needed to obtain consistent estimates of the optimal designs. For densely observed functional data, cross-sectional estimates are sufficient. Methods to overcome the difficulty of sparse sampling when targeting the mean and covariance functions have been addressed by various researchers (Yao *et al.*, 2005a, b; Staniswalis and Lee, 1998; Li and Hsing, 2010). The sparsely sampled case is different from the more commonly considered situation of densely sampled functional data, where individual curves can be consistently estimated by direct smoothing (Rice, 2004) and the covariance function is readily estimated by cross-sectional averaging (Ramsay and Silverman, 2005).

In the sparse case, these direct approaches do not lead to consistent estimates, because of the sparseness and lack of balance of the measurements. The way forward is to pool data for estimating mean and covariance functions, borrowing strength from the entire sample. For the required smoothing steps, we adopt one- and two-dimensional local linear smoothing, with further details in the following section. The estimated autocovariance function is further regularized by retaining only the positive eigenvalues and eigenvectors of the smoothed covariance function, so that $\hat{\Gamma}(s,t)$ is non-negative definite. An estimate $\hat{\sigma}^2$ of σ^2 is also needed and this is discussed in Section 5. For any p-dimensional vector $(t_1,\ldots,t_p)^{\mathrm{T}}$ of design points picked from the user-specified dense grid, we then have estimates $\hat{\boldsymbol{\mu}}_X$, $\hat{\boldsymbol{\Gamma}}_*$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\mathbf{C}}$ for $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}_*$, $\boldsymbol{\gamma}$ and $\mathbf{C}$ respectively. Then the estimated optimal designs are,

(a) for trajectory recovery,

$$\hat{\mathbf{t}}_X^* = \underset{(t_1,\ldots,t_p)^{\mathrm{T}}}{\arg\max} \int_{\mathcal{T}} \hat{\boldsymbol{\gamma}}^{\mathrm{T}}(t)\,\hat{\boldsymbol{\Gamma}}_*^{-1}\,\hat{\boldsymbol{\gamma}}(t)\,\mathrm{d}t, \tag{23}$$

where all integrals are implemented with trapezoidal integration,

(b) for scalar response regression,

$$\hat{\mathbf{t}}_Y^* = \underset{(t_1,\ldots,t_p)^{\mathrm{T}}}{\arg\max}\, \hat{\mathbf{C}}^{\mathrm{T}}\hat{\boldsymbol{\Gamma}}_*^{-1}\hat{\mathbf{C}}. \tag{24}$$

5. Numerical Implementation

5.1. Mean and covariance estimation via smoothing

First pooling sparse longitudinal data across subjects, we apply local linear estimators (Li and Hsing, 2010) to the resulting scatter plots, which depend on a bandwidth h as a tuning (smoothing) parameter. Writing $S_p\{t,(Q_j,V_j)_{j=1,\ldots,m},h\}$ for a local linear q-dimensional smoother (with $q=1$ or $q=2$) with output at the predictor level t and employing bandwidth h to smooth the scatterplot (Q_j,V_j), where $Q_j\in\mathcal{R}^q$, we obtain estimates $\hat{\mu}(t)$ for the mean function $\mu(t)$ as $S_1\{t,(t_{ij},U_{ij})_{i=1,\ldots,n,\,j=1,\ldots,m_i},h_\mu\}$, $t\in\mathcal{T}$, smoothing estimates $\tilde{\Gamma}(s,t)$ for the autocovariance function $\Gamma(s,t)$ as $S_2\{(s,t),(t_{ij},t_{ik},U_{ij}U_{ik})_{i=1,\ldots,n,\,j,k=1,\ldots,m_i,\,j\neq k},h_R\}-\hat{\mu}(s)\,\hat{\mu}(t), s,t\in\mathcal{T}$, and estimates $\hat{C}(t)$ for the cross-covariance function $C(t)$ as $S_1\{t,(t_{ij},Y_iU_{ij})_{i=1,\ldots,n,j=1,\ldots,m_i},h_S\}-\hat{\mu}(t)\,\hat{\mu}_Y, t\in\mathcal{T}$. We also obtain the estimate $\hat{\mu}_Y$ as the sample mean of the scalar responses.

Bandwidths for all smoothing steps are selected by cross-validation or generalized cross-validation. Details on the smoothing steps are in the on-line supplement section A.2, and assumptions for establishing consistency of the above smoothing steps in the longitudinal data context are provided in the on-line supplement section A.4. We used the function FPCA in ‘fdapace’, an R package for functional data analysis recently released on the Comprehensive R Archive Network, for smoothing and estimating the model components.

From the estimates $\tilde{\Gamma}(s,t)$ we then obtain estimates $\hat{\rho}_j$ and $\hat{\psi}_j$ for eigenvalues and eigenfunctions of predictor processes X by discretization and matrix spectral decomposition. The final autocovariance estimates $\hat{\Gamma}(s,t)$ are obtained by projecting on the space of non-negative and symmetric surfaces, simply by retaining only the positive eigenvalues and their corresponding eigenvectors (Hall *et al.*, 2008), yielding $\hat{\Gamma}(s,t)=\Sigma_{j=1,\,\hat{\rho}_j>0}^{K}\hat{\rho}_j\,\hat{\psi}_j(s)\,\hat{\psi}_j(t)$.

5.2. Stable covariance matrix inversion

As a practical implementation of the matrix inversion condition (9), we apply ridge regression (Hoerl and Kennard, 1970), enhancing the diagonal of the autocovariance surface $\hat{\Gamma}_*(s,t)$, as the matrices that need to be inverted are submatrices of this surface. Adding a suitable ridge parameter $\hat{\sigma}^2_{\mathrm{new}}$ at the diagonal ensures positive definiteness of all relevant $p\times p$ submatrices:

$$\hat{\Gamma}_*(s,t)=\hat{\Gamma}(s,t)+\sigma^2_{\mathrm{new}}\delta_{st}. \tag{25}$$

Here $\delta_{st}=1$ if and only if $s=t$. The optimization procedures in equations (23) and (24) are then implemented with $\hat{\Gamma}_*(s,t)$ or $\hat{\mathbf{\Gamma}}_*=\hat{\mathbf{\Gamma}}+\sigma^2_{\text{new}}I_p$. We explored two options to select the ridge parameter σ^2_{new}.

5.2.1. *Cross-validation*

For *trajectory recovery*, target criteria are the average root-mean-squared error ARE and the relative average root-mean-squared error ARE*, defined as

$$\text{ARE}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m_i}\sum_{j=1}^{m_i}[U_i(t_{ij})-\hat{B}_{-i}\{X_i(t_{ij})|\mathbf{U}\}]^2\right)^{1/2},$$

$$\text{ARE}^*=\sum_{i=1}^{n}\left(\frac{1}{m_i}\sum_{j=1}^{m_i}[U_i(t_{ij})-\hat{B}_{-i}\{X_i(t_{ij})|\mathbf{U}\}]^2\right)^{1/2}\Bigg/\sum_{i=1}^{n}\left\{\frac{1}{m_i}\sum_{j=1}^{m_i}U_i(t_{ij})^2\right\}^{1/2}, \qquad (26)$$

where $(U_i(t_{i1}),\ldots,U_i(t_{im_i}))^{\mathrm{T}}$ and $(\hat{B}_{-i}\{X_i(t_{i1})|\mathbf{U}\},\ldots,\hat{B}_{-i}\{X_i(t_{im_i})|\mathbf{U}\})^{\mathrm{T}}$ are observed measurements and plugged-in estimated best linear predictors for recovered processes with design $\mathbf{t}$ at the same time points. Leave-one-out versions of ARE and ARE* are easily obtained if we have densely measured functional data in the pilot study but are usually not viable when the pilot study consists of longitudinal data with sparse measurements, where observed subjects generally will not have recorded measurements at the selected optimal design points.

For *scalar response prediction*, the average prediction error APE and relative APE are natural criteria that can be easily cross-validated, irrespectively of the design of the pilot study,

$$\text{APE}=\left(\frac{1}{n}\sum_{i=1}^{n}[Y_i-\hat{B}_{-i}\{E(Y_i|X_i)|\mathbf{U}_i\}]^2\right)^{1/2},$$

$$\text{APE}^*=\left(\sum_{i=1}^{n}[Y_i-\hat{B}_{-i}\{E(Y_i|X_i)|\mathbf{U}_i\}]^2\right)^{1/2}\Bigg/\left(\sum_{i=1}^{n}Y_i^2\right)^{1/2} \qquad (27)$$

where $\hat{B}_{-i}\{E(Y_i|X_i)|\mathbf{U}_i\}$ is the estimated ith response obtained with estimated optimal designs that are obtained from a training sample leaving out the ith observation.

5.2.2. *Modified cross-validation*

A direct cross-validation approach is not feasible for the case of sparsely sampled pilot studies. In a modified approach, for each ridge parameter in a candidate set Ω, we repeatedly and randomly partition the training sample $\mathcal{S}$ from the sparsely sampled pilot study into two sets, $\mathcal{S}_{\mathrm{A}}$ and $\mathcal{S}_{\mathrm{B}}$, estimate model components from $\mathcal{S}_{\mathrm{A}}$ and then find the relatively best design $\mathbf{t}^*_{X,\mathcal{S}_{\mathrm{B}}}$ or $\mathbf{t}^*_{Y,\mathcal{S}_{\mathrm{B}}}$ by maximizing the criteria in equation (10) or (20) over $\mathcal{T}_{\mathrm{B}}$, which is the set of all available designs determined by the random configurations of the design points as observed for the subjects in the sample $\mathcal{S}_{\mathrm{B}}$. We then recover the trajectory or estimate the response for those subjects in $\mathcal{S}_{\mathrm{B}}$ where there is a match of the selected design $\mathbf{t}^*_{X,\mathcal{S}_{\mathrm{B}}}$ or $\mathbf{t}^*_{Y,\mathcal{S}_{\mathrm{B}}}$ with the design for that subject.

Combining L different random partitions, the mean ARE or APE is used to evaluate the performance of the ridge parameter choice in equation (25), yielding the selected parameter

$$\sigma^2_{\text{new}}=\underset{\Omega}{\arg\min}\sum_{l=1}^{L}\left\{\frac{1}{n_{\mathrm{B}}}\sum_{i_{\mathrm{B}}=1}^{n_{\mathrm{B}}}\left(\frac{1}{m_{i_{\mathrm{B}}}}\sum_{j=1}^{m_{i_{\mathrm{B}}}}[U_{i_{\mathrm{B}}}(t_{i_{\mathrm{B}},j})-\hat{B}_{\mathcal{S}_{\mathrm{A}}}\{X_{i_{\mathrm{B}}}(t_{i_{\mathrm{B}},j})|\mathbf{U}_{i_{\mathrm{B}}}\}]^2\right)^{1/2}\right\},$$

for trajectory prediction, and

$$\sigma^2_{\mathrm{new}} = \arg\min_{\Omega} \sum_{l=1}^{L} \left(\frac{1}{n_{\mathrm{B}}} \sum_{i_{\mathrm{B}}=1}^{n_{\mathrm{B}}} [Y_{i_{\mathrm{B}}} - \hat{B}_{\mathcal{S}_{\mathrm{A}}}\{E(Y_{i_{\mathrm{B}}}|X_{i_{\mathrm{B}}})|\mathbf{U}_{i_{\mathrm{B}}}\}]^2 \right)^{1/2}$$

for response prediction. Here, i_{B} is the index of the subjects in $\mathcal{S}_{\mathrm{B}}$ with available measurements at the selected optimal designs $\mathbf{t}^*_{X,\mathcal{S}_{\mathrm{B}}}$ and $\mathbf{t}^*_{Y,\mathcal{S}_{\mathrm{B}}}$ for trajectory recovery and prediction respectively, with n_{B} denoting the number of such subjects, and $m_{i_{\mathrm{B}}}$ is the number of measurements for the subject with index i_{B}. The estimator $\hat{B}_{\mathcal{S}_{\mathrm{A}}}$ is fitted on the basis of data from sample $\mathcal{S}_{\mathrm{A}}$ only, and index sets $\mathcal{S}_{\mathrm{A}}$ and $\mathcal{S}_{\mathrm{B}}$ depend on the random partition l. This method worked well in simulations and applications with longitudinal pilot studies.

5.3. *Sequential selection of design points*

Computationally, once the number of design points p has been specified, both trajectory recovery and scalar response prediction involve p-dimensional optimization. Exhaustive search over all combinations of grid points is very time consuming when employing optimization algorithms such as simulated annealing (Kirkpatrick *et al.*, 1983). A faster alternative is sequential selection, which is a greedy algorithm, where one searches for global optimal designs when $p=2$ as an initial step and then adds design points one by one iteratively, until the number of design points reaches the desired number. At each step, the target design point is the point that maximizes the selection criteria when adding it to the currently selected design points, which are carried forward unaltered. The sequential method is fast but does not guarantee finding the optimal solution. The performance differences of sequential and exhaustive search selection were found to be relatively small in simulation studies.

6. Simulation studies

We study the performance of optimal designs for trajectory recovery and scalar response prediction under two separate scenarios. In scenario 1 we consider the case where the pilot study generates densely observed functional data, and in scenario 2 the case where it generates sparse longitudinal data. Random trajectories are generated as $X_i(t) = \mu(t) + \Sigma_{k=1}^{K} \zeta_{ik}\, \psi_k(t)$, and observed data as $U_i(t_{ij}) = X_i(t_{ij}) + e_{ij}$, where $e_{ij} \sim N(0, 0.25)$.

For the regression case, the response is chosen as $Y_i = \int \beta(t)\, X_i(t)\, \mathrm{d}t + \epsilon_i = \zeta_{i,1} - 2\zeta_{i,2} + \zeta_{i,3} - 2\zeta_{i,4} + \epsilon_i$, i.e. as a linear combination of functional principal components of the process, where $\epsilon_i \sim N(0, 0.25)$ are independently and identically distributed. We specify $\mathcal{T} = [0, 10]$, mean function $\mu(t) = \frac{1}{2}t^2 + 2\sin(t) + 3\cos(2t)$, $t \in \mathcal{T}$, and include 10 functional principal components, with eigenvalues $(30, 20, 12, 8, 30/25, 30/36, 30/49, 30/64, 30/81, 30/100)^{\mathrm{T}}$, and corresponding eigenfunctions $\psi_k(t) = \surd(2/10)\cos\{(k/10)\pi t\}$, for $k = 1, 2, \ldots, 10$, and generate data for 100 subjects in the training sample and 1000 in the testing sample. The measurement locations t_{ij} are assumed to form a dense grid for simulation scenario 1 and are sparse, with a random number of 4–8 measurement locations per subject for scenario 2. Fig. 9 in the on-line supplement section A.9 shows the spaghetti plot of the data for the subjects in one training sample.

For both trajectory recovery and prediction for functional linear regression, we applied the proposed procedures for the subjects in the training sample to construct the optimal designs for $p = 2, 3, 4$ with exhaustive search and $p = 2, 3, \ldots, 8$ with sequential search. Each simulation scenario was repeated 100 times. The optimal ridge parameter σ^2_{new} in equation (25) was determined by cross-validation for scenario 1 and modified cross-validation for scenario 2 (see Section 5.2). We compare the median performance with regard to (relative) average root-squared error

Table 1. Comparing optimal with random designs in 100 simulations, in terms of mean ARE and relative ARE* (26) (in parentheses) for trajectory recovery and APE and relative APE* (27) (in parentheses) for response prediction in a functional linear model†

Number of design points, p	*Design*	*Mean ARE (ARE*)*		*Mean APE (APE*)*	
		Dense	*Sparse*	*Dense*	*Sparse*
2	Optimal	1.74 (0.091)	1.84 (0.102)	3.37 (0.272)	9.41 (0.759)
	Random	1.91 (0.101)	2.12 (0.117)	10.65 (0.870)	10.79 (0.906)
3	Optimal	1.37 (0.072)	1.59 (0.088)	2.79 (0.224)	7.63 (0.613)
	Random	1.65 (0.087)	1.88 (0.104)	8.56 (0.687)	9.80 (0.793)
4	Optimal	1.02 (0.054)	1.34 (0.074)	2.54 (0.204)	6.87 (0.553)
	Random	1.46 (0.077)	1.74 (0.096)	6.44 (0.499)	7.65 (0.616)

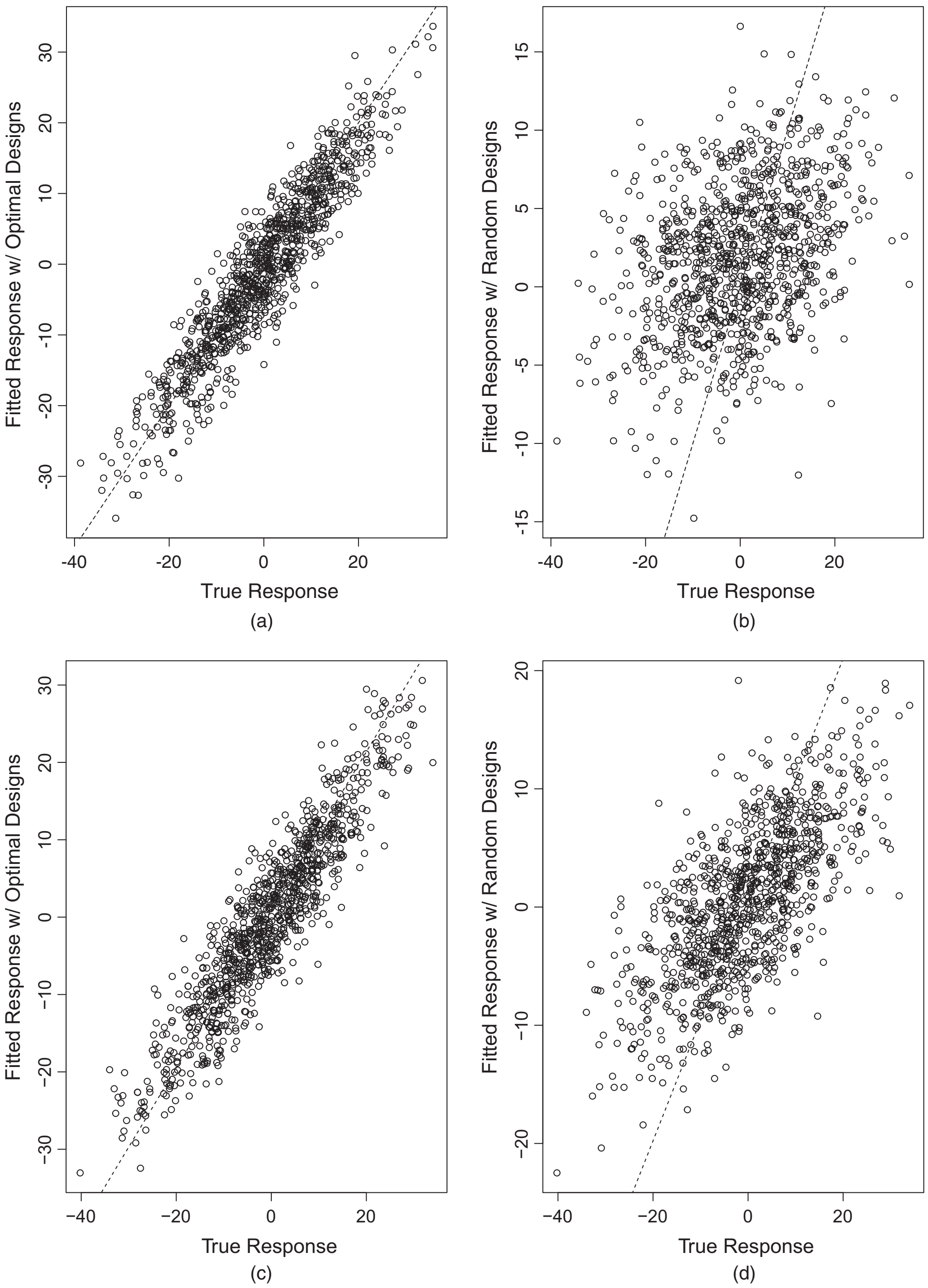
†Exhaustive search was used.

ARE and (relative) average prediction error APE defined in equations (26) and (27) for optimal designs and random designs. These random designs use the same number of design points as the estimated optimal designs; however, the locations are sampled from a uniform distribution over all possible locations (we provide further comments on the rationale of comparing with random designs in the on-line supplement section A.5). The results are summarized in Table 1 and are illustrated in Fig. 2 (and also Fig. 10 in the on-line supplement section A.9) for the case where the pilot study is longitudinal with sparsely sampled functional data.

These simulations demonstrate that the proposed optimal designs exhibit better performance than random designs for both trajectory recovery and scalar response prediction, especially for sparse functional data. For trajectory recovery, Fig. 10 (in the on-line supplement section A.9) for sparse pilot designs indicates that recovered trajectories obtained from optimal designs are closer to the underlying true curves than those obtained from median performance random designs. For scalar response prediction, Fig. 2 corroborates the results from Table 1, namely that optimal designs outperform median performance random designs for the prediction of a subject's response from a few observed measurements only.

The boxplots of ARE and APE for 100 simulation runs in Fig. 3 and Fig. 10 (in the on-line supplement section A.9) visualize the variation of performance over the simulations, comparing optimal and median performance random designs and also show the improvement in performance as the number of design points p increases. For densely observed functional data, various penalization schemes are possible to select p, minimizing the sum of ARE (or APE) and a penalty that increases with increasing p. However, such schemes are not directly applicable to longitudinal data, because subjects rarely will have been observed at the selected design points for trajectory recovery or prediction. In practice, an upper bound for p often will be dictated by cost. Simulation results for the effect of ridge parameter selection on the performance of optimal designs showed that the selection proposed works well (see the on-line supplement section A.7).

To summarize the simulation results, optimal designs performed very well and in any case better than random designs for both trajectory recovery and scalar response prediction. The costs that are incurred when adopting the much faster sequential search algorithm are quite small. Unsurprisingly, the performance of the optimal designs was seen to improve with increasing number of design points.



(a)

(b)

(c)

(d)

Fig. 2. Simulation results for sparsely sampled pilot data: (a), (b) scatter plots for fitted *versus* true responses for predicting a scalar response based on three design points; (c), (d) scatter plots for fitted *versus* true responses for predicting a scalar response based on four design points; (a), (c) results for optimal designs; (b), (d) results for random designs with median performance

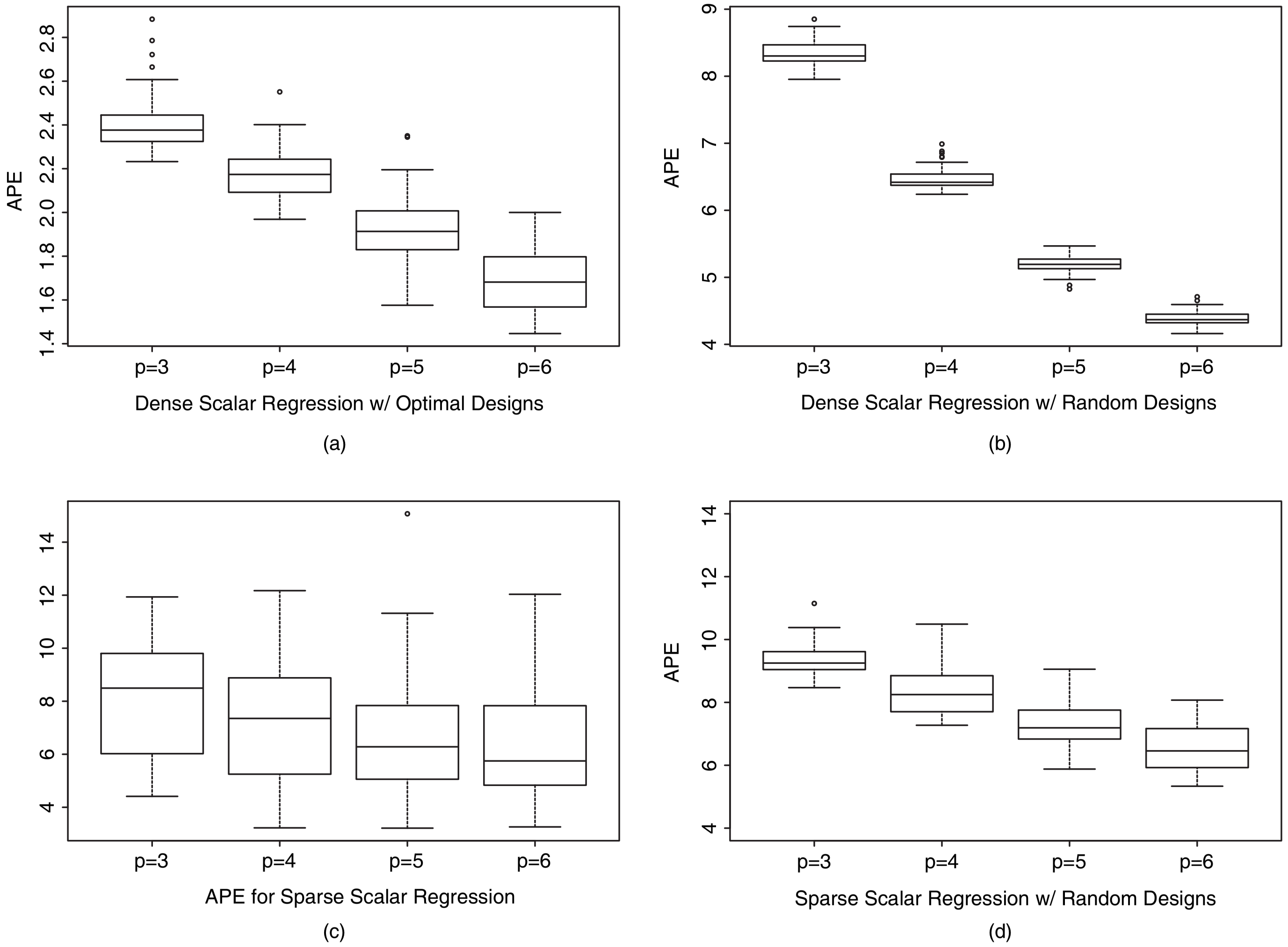


Fig. 3. Boxplots of APE from 100 simulation runs for scalar response regression for both (a), (b) dense and (c), (d) sparse scenarios with the sequential search algorithm, with $p = 3, 4, 5, 6$: (a), (c) results for the optimal design; (b), (d) results for the median performance random design

7. Data illustrations

7.1. Mediterranean fruit fly egg laying

The Mediterranean fruit fly data, which were described in Carey *et al*. (2002), consist of egg laying profiles for 1000 female Mediterranean fruit flies. For each fly, daily measurements on the number of eggs laid during the day are available from birth to death. A biologically relevant regression problem is to utilize the partial egg laying profile from day 1 to day 30 to predict the number of eggs that will be laid during the remaining lifetime for each fly. This yields information about the reproductive potential of the fly at age 30 days, which is related to its evolutionary fitness (Kouloussis *et al*., 2011).

We aim to find the optimal design points for this scalar response regression problem. To prevent censoring, we include only flies that live beyond 30 days. The measurements in the pilot data are dense and regular. Since daily egg laying counts require constant monitoring of the flies, reducing this task to monitoring the flies at a few time points is useful to scale up such studies. It is of additional interest to identify key days that are relevant for the prediction of the egg laying potential. We use the complete available egg laying profiles to find optimal design points that are most relevant for the prediction of the remaining total number of eggs via a functional linear regression model.

From the 667 subjects surviving more than 30 days, we select a training sample of 500 flies, and a testing sample that consists of the remaining 167 flies. Fig. 4 shows the spaghetti plot for a subset of the training sample. We apply the proposed methodology to find optimal designs for $p=3$. The relationship between observed and predicted responses is shown in Fig. 5. The relative APE (27) is 0.483 when using optimal designs, as opposed to 0.665 by using random designs with median performance for $p=3$. Fig. 5 provides a graphical illustration that optimal designs clearly

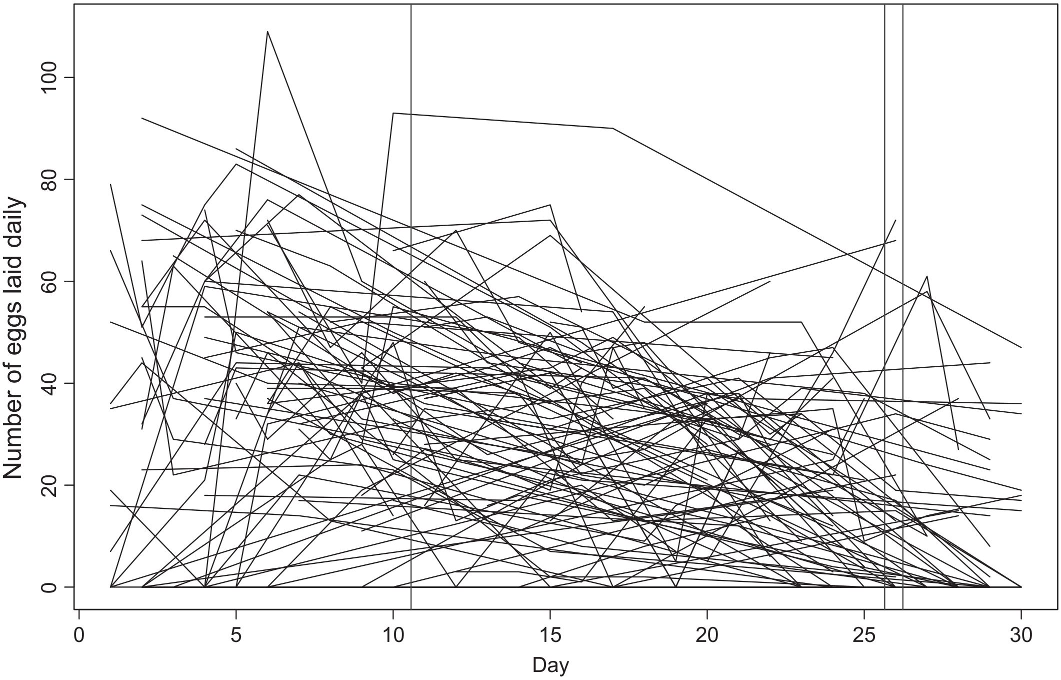


Fig. 4. Spaghetti plot for a subset of the training sample for predicting egg laying potential for female medflies from their egg laying profiles: |, locations of the optimal design points for p = 3

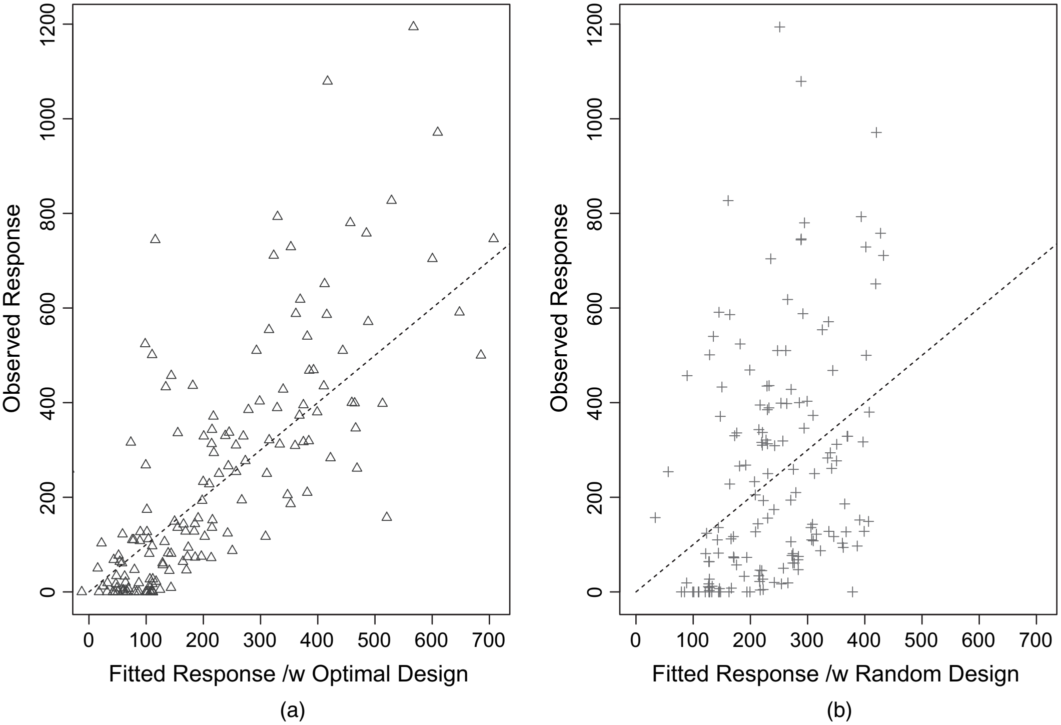


Fig. 5. Observed *versus* fitted (predicted) numbers of eggs that female medflies lay in their remaining lifetime when using $p = 3$ design points in a functional linear regression model (⋯⋯, 45° line): (a) predictions for optimal designs; (b) predictions for median performance random designs

outperform randomly chosen designs with median performance. The three selected optimal design points for $p = 3$ are at days 10, 25 and 26. Their locations are shown as vertical lines in Fig. 4. These design point locations are consistent with previous findings that both the intensity of egg laying at earlier ages and the rate of decline at older ages are closely related to the reproductive potential for individual flies (Müller *et al*., 2001). Two design points are selected on consecutive days 25 and 26, indicating that these locations are useful to gauge the rate of decline in egg laying, which corresponds to quantifying a derivative in this age range.

7.2. The Baltimore Longitudinal Study of Aging

For the BLSA we aim at identifying optimal designs for recovering BMI profiles from sparse measurements and for predicting a subject's systolic blood pressure (SBP) at old age. To construct optimal designs, we use available pilot data that come from the longitudinal BLSA study, where measurements of BMI are sparse and irregularly spaced. To avoid bias due to censoring effects, we include only subjects with non-missing date of death and for whom age at death is above 70 years and consider only the available measurements that were taken within the age range from 45 to 70 years. We also exclude subjects who had fewer than four measurements. The response is taken to be the last SBP measurement before death. Within the study sample, 496 subjects met these criteria and were included in the analysis. A subset of the data, where measurements are connected by straight lines, is shown in Fig. 1.

A decisive difference between these data and the previous data illustration is that the BMI trajectories that are assumed to generate the observed data are not available, as these pilot data

Table 2. Optimal designs (ages where measurements are placed) and associated coefficients of determination (in parentheses) for different numbers of design points for the BLSA data, for recovering BMI and for predicting SBP

p	*Designs (years) for BMI trajectory recovery* (R_X^2)	*Designs (years) for SBP prediction* (R_Y^2)
3	46.5, 58, 68 (0.951)	45, 45.5, 53 (0.863)
4	46, 56.5, 62.5, 70 (0.973)	45, 45.5, 52, 52.5 (0.874)
5	47, 47.5, 60.5, 61.5, 70 (0.989)	45, 45.5, 53, 53.5, 68.5 (0.886)

are from a typical longitudinal study with inherently sparse and irregular measurement times. Therefore, only indirect information is available for each subject about the underlying trajectory and only conditional inference about the trajectories is possible, conditioning on the available measurements for each subject (Yao *et al.*, 2005a). This is the typical situation that we face when constructing optimal designs in the common situation where the available pilot data are from a longitudinal study. Nevertheless, the construction of optimal designs is still possible since they depend only on the covariance structure of the data, which can still be consistently estimated. To implement the proposed optimal designs for this situation, we use the modified cross-validation method to select the ridge parameter.

Here the construction of optimal designs is intended for future data collection in longitudinal studies that will adopt the same fixed optimal design for the subjects to be included in the subsequent study. Because the pilot data are coming from a longitudinal study with random measurement locations, subjects who are included in the pilot study will normally not have measurements at the optimal design locations. As in addition their actual functional trajectories are unknown, it is not possible within the framework of the BLSA study to evaluate directly the performance of the optimal designs in terms of recovering the BMI trajectories or predicting old age SBP. Alternatively, we consider performance as measured by the coefficients of determination in expressions (8) and (19). The optimal designs selected, along with their associated coefficients of determination, are listed in Table 2 for $p=3,4,5$, for both prediction and trajectory recovery, where exhaustive search was used to determine these designs.

The following findings are of interest for this and similar situations where we have longitudinal pilot data: first, the designs for the prediction and the recovery task were found to differ somewhat, especially in terms of older age measurements that are part of the optimal designs for trajectory recovery if $3 \leqslant p \leqslant 4$ but are somewhat less relevant for predicting the SBP response at 70 years. Therefore, optimal designs for old age SBP prediction from BMI will not require measurements at older ages. Second, the constructed optimal designs for the same target quantity are consistent in the sense that, as p increases, additional design points are added while the previously selected design points are still viable so that selected optimal designs to some extent form a nested sequence, even when exhaustive search is used to determine these designs. Third, the design points for trajectory recovery tend to be more evenly distributed than those for response prediction. In other words, design points that are more or less uniformly distributed across the age range seem to achieve the goal of trajectory recovery well.

The latter point was also seen in the bicycle sharing data, which are described in the on-line supplement section A.6. The reason behind this might be that the random variation in trends or curvature across subjects is relatively uniform in these examples (see Fig. 1). The optimal design points for prediction of old age SBP are generally clustering around earlier ages, which is potentially of interest for public health and prevention. Finally, the coefficients of determination for both trajectory recovery and prediction are increasing with p, showing better performance with increasing number of design points, as was also seen in the simulations. To select the optimal designs together with the optimal p, one can use a suitable penalty for larger p that might reflect the data collection cost in specific applications.

Specifically for the BLSA study, comparing designs with different p, a general recommendation would be to take the first measurement at around age 46 or 47 years, then the second within the age range 58–62 years and finally a third at around age 70 years if $p=3$ and BMI trajectory recovery is the objective. For predicting the SBP, optimal designs are more concentrated in the first half of the age range, and it would be sufficient to take measurements at age 45, 45.5 and around age 53 years.

8. Theory

We establish asymptotic consistency and rates of convergence for the optimal designs proposed. Assumptions (A1)–(A7) and the proof of the following main result are in the on-line supplement section A.4. The rates of convergence that we report here are for the case that the pilot study is longitudinal and generates sparse functional or longitudinal data. In what follows, h_μ, h_S and h_R are bandwidths for local linear smoothers for the mean function, cross-covariance function and autocovariance function, as specified in expressions (32), (33) and (34) in the on-line supplement section A.2.

Theorem 1. Assume that assumptions (A1)–(A7) in the on-line supplement hold, and that criteria R^2_X and R^2_Y are locally concave around the optimal design $\mathbf{t}^*_X$ for trajectory recovery (10) and $\mathbf{t}^*_Y$ for scalar response regression (20) respectively, where $\mathbf{t}^*_X$, $\mathbf{t}^*_Y \in \mathcal{T}(\delta_0)$, defined at equation (9). For $\theta_n = h_R^2 + \{\log(n)/(nh_R^2)\}^{1/2}$ and, given fixed p, the estimated optimal designs for trajectory recovery $\hat{\mathbf{t}}^*_X$ given by equation (10) and for scalar response regression $\hat{\mathbf{t}}^*_Y$ given by equation (20) satisfy

$$\|\hat{\mathbf{t}}^*_X - \mathbf{t}^*_X\|_p = O(\theta_n^{1/2}) \qquad \text{almost surely as } n \to \infty$$

and

$$\|\hat{\mathbf{t}}^*_Y - \mathbf{t}^*_Y\|_p = O(\theta_n^{1/2}) \qquad \text{almost surely as } n \to \infty$$

where $\|\mathbf{t}-\mathbf{s}\|_p = \max_{1\leqslant j\leqslant p} |\mathbf{t}_{(j)} - \mathbf{s}_{(j)}|$, where $\mathbf{t}_{(j)}$ and $\mathbf{s}_{(j)}$ are the jth-order statistics of designs $\mathbf{t}$ and $\mathbf{s}$. Here, we assume in addition that the smoothing bandwidths satisfy $h_R^2 \lesssim h_\mu \lesssim h_R$, and $h_\mu \sim h_S$, where $a_n \lesssim b_n$ means that $a_n = O(b_n)$, and $a_n \sim b_n$ means that $c \leqslant a_n/b_n \leqslant C$ for constants $0<c<C<\infty$.

This result demonstrates the consistency of the estimated optimal designs, including rates of convergence to the true optimal designs. This provides a theoretical justification for our methods. The rate of convergence θ_n is determined by the rate at which the bandwidth h_R for smoothing the cross-covariance surface of the underlying smooth stochastic process converges to 0.

Theorem 1 is proved by first establishing uniform convergence of the optimization criteria (10) and (20) with plugged-in estimators of the auxiliary quantities. The second step then is to prove the convergence of the estimated maximizer to the true maximizer. If the pilot study has dense and regular designs, we can apply cross-sectional covariance and mean estimation and by similar arguments to those provided here obtain analogous results as in theorem 1 with the faster rate $\theta_n = n^{-1/2}$.

9. Discussion and concluding remarks

We have developed a new method to obtain optimal designs for longitudinal data, by considering such data to be instances of functional data, i.e. by assuming that they are generated from an underlying (but unobservable) smooth stochastic process. This perspective makes it possible to take a pilot sample of sparsely observed longitudinal data and to construct optimal design points for future observations, where optimality refers to various criteria and loss functions.

For the trajectory recovery task, the optimal designs are related to the shapes of the first few eigenfunctions that explain most of the variation. Specifically, the optimal design points are likely to cluster around areas where the variation in the underlying process across subjects is large, and that might correspond to areas where the first few eigenfunctions have peaks or valleys (Hall and Vial, 2006), or generally areas where they are large in absolute value. In contrast, for the scalar regression task, the optimal designs cannot be easily understood from the shape of the eigenfunctions, because these designs depend not only on the autocovariance, but equally on the cross-covariance between responses and predictor trajectories.

Simulations and data analyses show that optimal designs can lead to substantial gains over random designs for both trajectory recovery and prediction. The optimal designs constructed thus provide guidance for the planning and collection of longitudinal data. The method proposed is conceptually straightforward and the estimating procedure is easy to implement. The method can be applied to a wide range of studies where dense measurements of underlying trajectories would be desirable but cannot be obtained because of various constraints. In various engineering and medical applications, dense measurement designs are too expensive to obtain and therefore sparse designs need to be used. Some further extensions are discussed in the on-line supplement section A.8.

Although our focus in this paper is on optimal trajectory recovery and optimal prediction as two pertinent and important criteria, there are many other conceivable targets, such as predicting a functional response, or specific features of the underlying random trajectories, e.g. integrals or derivatives. It is also possible to select a mixed target criterion, targeting both trajectory recovery and optimal prediction. One can then apply the same methodology as we describe here to find the optimal designs for such a mixed target criterion that blends various objectives, with relative weights given to each. For any given design, we can use the proposed criteria to evaluate its relative suitability for both prediction of a response as well as obtaining trajectory estimates in comparison with competing designs, e.g. by comparing the relative values of ARE (26) and APE (27). This enables evaluation of the relative merits of potentially suboptimal designs that sometimes may be more convenient than optimal designs or may be the result of extraneous constraints in data collection.

Our methods are also applicable in situations where in principle one can sample functional data densely but in future data collection plans to sample at only a few key locations. We find that the construction of optimal designs for longitudinal studies benefits in many ways from the adoption of a functional approach.

Acknowledgement

This research was supported by National Science Foundation grants DMS-1104426 and DMS-1407852.

References

Anisimov, V. V., Fedorov, V. V. and Leonov, S. L. (2007) Optimal design of pharmacokinetic studies described by stochastic differential equations. In *Advances in Model-oriented Design and Analysis* (eds J. Lopez-Fidalgo, J. M. Rodríguez-Díaz and B. Torsney), pp. 9–16. New York: Springer.

Cardot, H., Ferraty, F. and Sarda, P. (1999) Functional linear model. *Statist. Probab. Lett.*, **45**, 11–22.

Carey, J. R., Liedo, P., Harshman, L., Zhang, Y., Müller, H.-G., Partridge, L. and Wang, J.-L. (2002) Life history response of Mediterranean fruit flies to dietary restriction. *Agng Cell*, **1**, 140–148.

Delaigle, A., Hall, P. and Bathia, N. (2012) Componentwise classification and clustering of functional data. *Biometrika*, **99**, 299–313.

Fedorov, V. V. and Leonov, S. L. (2013) *Optimal Design for Nonlinear Response Models*. Boca Raton: CRC Press.

Ferraty, F., Hall, P. and Vieu, P. (2010) Most-predictive design points for functional data predictors. *Biometrika*, **97**, 807–824.

Guo, W. (2004) Functional data analysis in longitudinal settings using smoothing splines. *Statist. Meth. Med. Res.*, **13**, 49–62.

Hall, P., Müller, H.-G. and Yao, F. (2008) Modelling sparse generalized longitudinal observations with latent Gaussian processes. *J. R. Statist. Soc.* B, **70**, 703–723.

Hall, P. and Vial, C. (2006) Assessing extrema of empirical principal component functions. *Ann. Statist.*, **34**, 1518–1544.

He, G., Müller, H. and Wang, J. (2000) Extending correlation and regression from multivariate to functional data. In *Asymptotics in Statistics and Probability* (ed. M. L. Puri), pp. 301–315. Leiden: VSP International.

Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Kouloussis, N. A., Papadopoulos, N. T., Katsoyannos, B. I., Müller, H.-G., Wang, J.-L., Su, Y.-R., Molleman, F. and Carey, J. R. (2011) Seasonal trends in ceratitis capitata reproductive potential derived from live-caught females in Greece. *Entmol. Experim. Appl.*, **140**, 181–188.

Li, Y. and Hsing, T. (2010) Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.*, **38**, 3321–3351.

McKeague, I. W. and Sen, B. (2010) Fractals with point impact in functional linear regression. *Ann. Statist.*, **38**, 25–59.

Mentre, F., Mallet, A. and Baccar, D. (1997) Optimal design in random-effects regression models. *Biometrika*, **84**, 429–442.

Müller, H.-G., Carey, J. R., Wu, D., Liedo, P. and Vaupel, J. W. (2001) Reproductive potential predicts longevity of female Mediterranean fruit flies. *Proc. R. Soc. Lond.* B, **268**, 445–450.

Pearson, J. D., Morrell, C. H., Brant, L. J., Landis, P. K. and Fleg, J. L. (1997) Age-associated changes in blood pressure in a longitudinal study of healthy men and women. *J. Gerontol.* A, **52**, M177–M183.

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

Rice, J. A. (2004) Functional and longitudinal data analysis: perspectives on smoothing. *Statist. Sin.*, **14**, 631–648.

Rice, J. A. and Wu, C. O. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253–259.

Shock, N. W., Greulich, R. C., Costa, P. T., Andres, R., Lakatta, E. G., Arenberg, D. and Tobin, J. D. (1984) Normal human aging: the Baltimore Longitudinal Study of Aging. National Institutes of Health.

Staniswalis, J. G. and Lee, J. J. (1998) Nonparametric regression analysis of longitudinal data. *J. Am. Statist. Ass.*, **93**, 1403–1418.

Xiang, D., Qiu, P. and Pu, X. (2013) Nonparametric regression analysis of multivariate longitudinal data. *Statist. Sin.*, **23**, 769–789.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005a) Functional data analysis for sparse longitudinal data. *J. Am. Statist. Ass.*, **100**, 577–590.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005b) Functional linear regression analysis for longitudinal data. *Ann. Statist.*, **33**, 2873–2903.

Zagoraiou, M. and Baldi Antognini, A. (2009) Optimal designs for parameter estimation of the Ornstein–Uhlenbeck process. *Appl. Stoch. Modls Bus. Indstry*, **25**, 583–600.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Online supplement’.