*Biometrika* (2010), **97**, 1, *pp.* 49–64
© 2010 Biometrika Trust
*Printed in Great Britain*

doi: 10.1093/biomet/asp069
Advance Access publication 18 January 2010

# Functional quadratic regression

BY FANG YAO

*Department of Statistics, University of Toronto, 100 Saint George Street, Toronto, Ontario M5S 3G3, Canada*

fyao@utstat.toronto.edu

AND HANS-GEORG MÜLLER

*Department of Statistics, University of California, Davis, One Shields Avenue, Davis, California 95616, U.S.A.*

mueller@wald.ucdavis.edu

## SUMMARY

We extend the common linear functional regression model to the case where the dependency of a scalar response on a functional predictor is of polynomial rather than linear nature. Focusing on the quadratic case, we demonstrate the usefulness of the polynomial functional regression model, which encompasses linear functional regression as a special case. Our approach works under mild conditions for the case of densely spaced observations and also can be extended to the important practical situation where the functional predictors are derived from sparse and irregular measurements, as is the case in many longitudinal studies. A key observation is the equivalence of the functional polynomial model with a regression model that is a polynomial of the same order in the functional principal component scores of the predictor processes. Theoretical analysis as well as practical implementations are based on this equivalence and on basis representations of predictor processes. We also obtain an explicit representation of the regression surface that defines quadratic functional regression and provide functional asymptotic results for an increasing number of model components as the number of subjects in the study increases. The improvements that can be gained by adopting quadratic as compared to linear functional regression are illustrated with a case study that includes absorption spectra as functional predictors.

*Some key words*: Absorption spectra; Asymptotics; Functional data analysis; Polynomial regression; Prediction; Principal component.

## 1. INTRODUCTION

Data that include a functional predictor in the form of a smooth random trajectory are increasingly common. Typical scenarios in which such data arise include frequently monitored trajectories such as in movement tracking (Faraway, 1997) and longitudinal studies where the trajectory is probed through noisy and often sparse and irregularly spaced measurements. Regression models that can handle functional predictors are therefore needed for a variety of settings and applications, and many aspects of these functional regression relations remain open problems. We consider here the case where a functional predictor is paired with a scalar response. Examples of such situations include various biological trajectories with regular observations as predictors (Kirkpatrick & Heckman, 1989) and are also commonly encountered in biodemographic applications (Müller & Zhang, 2005). In many longitudinal studies, measurements of longitudinal

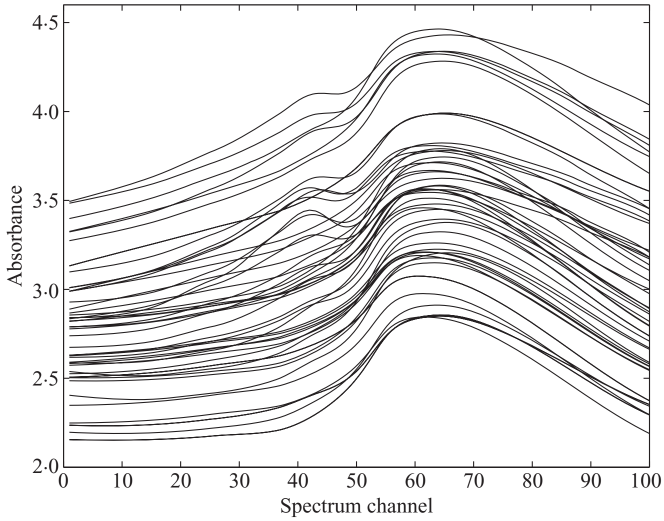


Fig. 1. The functional predictor trajectories, consisting of 100 channel spectra of log-transformed absorbances, for 50 randomly selected meat specimens.

trajectories are recorded only intermittently and often at time-points that do not conform with a regular grid (Müller, 2005).

As an example of a typical functional regression problem, consider the sample of trajectories in Fig. 1. Displayed are 50 randomly selected 100-channel absorption spectra that are used to predict the composition of food samples. The spectra shown are from meat specimens and the goal is to predict fat contents. These spectra have been densely sampled at 100 support points and are seen to be quite smooth. Taking advantage of the smoothness of these trajectories is key to the efficient modelling of regression relationships that include functional predictors. The functional nature of the predictors and the measurements need to be adequately reflected in the statistical modelling of such data (Rice, 2004; Zhao et al., 2004).

In this paper, we consider both densely and sparsely sampled longitudinal predictor data. In many longitudinal studies for which one wishes to apply functional regression, the predictor process must be inferred from noisy and sparse measurements (James et al., 2000; Yao et al., 2005b). For the most appropriate analysis of a given set of data with functional predictors, one desires a variety of readily available models, from which the data analyst can choose the most appropriate approach. The situation is analogous to the case of ordinary regression models involving ve2004 data, where quadratic models are commonplace and vcup207(seen)--2itudin-296(stn-29ongitud the extension of the linear model to the case of a polynomial functional relationship, analogous to the extension of linear regression to polynomial regression in traditional regression settings and highlight the important special case of a quadratic regression.

To achieve the regularization that is necessary for any functional regression model, we project predictor processes on a suitable basis of the underlying function space, which is then truncated at a reasonable number of included components. We implement this regularization through the eigenbasis of the predictor processes, which leads to parsimonious representations. A key finding is that the functional polynomial regression model can be represented as a polynomial regression model in the functional principal component scores of the predictor process. Accordingly, it is convenient to implement the model as polynomial regression in the principal components of predictor processes. The representation in terms of functional principal components makes it possible to include both densely and sparsely observed predictor trajectories, allows for simple numerical implementation, and enables us to obtain asymptotic consistency results within the framework of a general measurement model.

## 2. Functional linear and polynomial regression

### 2·1. *From functional linear to quadratic regression*

The functional regression models we consider include a functional predictor paired with a scalar response. The predictor process is assumed to be square integrable and is defined on a finite domain $\mathcal{T}$, with mean function $E\{X(t)\} = \mu_X(t)$ and covariance function $\operatorname{cov}\{X(s), X(t)\} = G(s, t)$ for $s, t \in \mathcal{T}$. The covariance function $G$ can be decomposed by means of the eigenvalues and eigenfunctions of the autocovariance operator of $X$. Denoting eigenvalue/eigenfunction pairs by $\{(\lambda_1, \phi_1), (\lambda_2, \phi_2), \ldots\}$, ordered according to $\lambda_1 \geqslant \lambda_2 \geqslant \cdots$, one obtains $G(s, t) = \sum_k \lambda_k \phi_k(s)\phi_k(t)$. The well-established linear functional regression model with scalar response (Ramsay & Dalzell, 1991) is given by

$$E(Y \mid X) = \mu_Y + \int_{\mathcal{T}} \beta(s) X^c(s)\, ds, \tag{1}$$

where $X^c(t) = X(t) - \mu_X(t)$ denotes the centered predictor process. The regression parameter function $\beta$ is assumed to be smooth and square integrable.

To estimate the function $\beta$, some form of regularization is needed, for which we employ truncated basis representations of predictor processes $X$. We choose the orthonormal eigenfunctions of predictor processes $X$ as the basis; alternative choices such as the wavelet basis may prove convenient in some applications (Morris & Carroll, 2006). When selecting the eigenbasis, one takes advantage of the equivalence between the predictor process and the countable sequence of uncorrelated functional principal components. These scores are the random coefficients $\xi_k$ in the Karhunen–Loève representation

$$X(t) = \mu_X(t) + \sum_k \xi_k \phi_k(t), \quad t \in \mathcal{T}, \tag{2}$$

with $\xi_k = \int X^c(t)\phi_k(t)\, dt$. They are uncorrelated with zero mean and $\operatorname{var}(\xi_k) = \lambda_k$.

While the functional linear model in (1) has been well investigated and has proven useful in many applications, it is desirable to develop a class of more general parametric functional regression models for situations where the functional linear model is inadequate. If a functional linear model does not provide an appropriate fit, a natural alternative is to move from a linear to a quadratic functional regression model, similarly to the situation in ordinary regression. This approach follows the classical strategy to embed an ill-fitting model into a larger class of models. It is thus natural to consider a quadratic regression relationship when moving one step beyond the functional linear model, or on occasion a functional relation that involves a polynomial of order higher than 2. The functional linear model is always included as a special case.

The quadratic functional regression relationship involves a square integrable univariate linear parameter function $\beta(t)$ and a square integrable bivariate quadratic parameter function $\gamma(s, t)$, and is given by

$$E(Y \mid X) = \alpha + \int_{\mathcal{T}} \beta(t) X^c(t)\, dt + \int_{\mathcal{T}} \int_{\mathcal{T}} \gamma(s, t) X^c(s) X^c(t)\, ds dt, \tag{3}$$

where $\alpha$ is an intercept. The linear part is seen to be the same as that in model (1), while a quadratic term has been added. This term reflects that beyond the effect that the ensemble of the values of $X^c(t)$, $t \in \mathcal{T}$, has on the response, the products $\{X^c(s)X^c(t)\}$, $s, t \in \mathcal{T}$, and in particular the square terms $\{X^c(t)\}^2$, $t \in \mathcal{T}$, are included as additional predictors.

Since the eigenfunctions $\{\phi_k\}_{k=1,2,\ldots}$ of the process $X$ form a complete basis, the regression parameter functions in (3) can be represented using this basis,

$$\beta(t) = \sum_{k=1}^{\infty} \beta_k \phi_k(t), \qquad \gamma(s, t) = \sum_{k,\ell=1}^{\infty} \tilde{\gamma}_{k\ell} \phi_k(s) \phi_\ell(t), \tag{4}$$

for suitable sequences $(\beta_k)_{k=1,2,\ldots}$ and $(\tilde{\gamma}_{k\ell})_{k,\ell=1,2,\ldots}$ with $\sum_k \beta_k^2 < \infty$ and $\sum_{k,\ell} \tilde{\gamma}_{k\ell}^2 < \infty$. Substituting representations (2) and (4) for the components in the quadratic model (3) and applying the orthonormality property of the eigenfunctions, one finds that the functional quadratic model in (3) can be alternatively expressed as a function of the scores $\xi_k$ of predictor processes $X$,

$$E(Y \mid X) = \alpha + \sum_{k=1}^{\infty} \beta_k \xi_k + \sum_{k=1}^{\infty} \sum_{\ell=1}^{k} \gamma_{k\ell} \xi_k \xi_\ell, \tag{5}$$

where $\gamma_{k\ell} = 2\tilde{\gamma}_{k\ell}$ for $k \neq \ell$ and $\gamma_{k\ell} = \tilde{\gamma}_{k\ell}$ for $k = \ell$. We also note that model (3) implies the constraint $\mu_Y = E(Y) = \alpha + \sum_k \gamma_{kk} \lambda_k$, i.e. the intercept has the representation $\alpha = \mu_Y - \sum_k \gamma_{kk} \lambda_k$.

### 2·2. *Functional polynomial regression*

Considering the more general case of a polynomial regression, we define the $p$th-order ($p \geqslant 3$) functional polynomial model in analogy to (3) as follows:

$$\begin{aligned} E(Y \mid X) = \alpha &+ \int_{\mathcal{T}} \beta(t) X^c(t)\, dt + \int_{\mathcal{T}^2} \gamma(s, t) X^c(s) X^c(t)\, ds dt \\ &+ \int_{\mathcal{T}^3} \gamma_3(t_1, t_2, t_3) X^c(t_1) X^c(t_2) X^c(t_3)\, dt_1 dt_2 dt_3 \\ &+ \int_{\mathcal{T}^p} \gamma_p(t_1, \ldots, t_p) X^c(t_1) \ldots X^c(t_p) dt_1 \ldots dt_p, \end{aligned}$$

where again $\alpha$ is the intercept and $\beta, \gamma, \gamma_j$ ($j = 3, \ldots, p$) are the linear, quadratic and $j$th-order regression parameter functions, defining the effects of the corresponding interactions. Using the same arguments as those leading to (5), this model can also be written in terms of the predictor functional principal components,

$$\begin{aligned} E(Y \mid X) = \alpha &+ \sum_{j_1 \geqslant 1} \beta_{j_1} \xi_{j_1} + \sum_{j_1 \leqslant j_2} \gamma_{j_1 j_2} \xi_{j_1} \xi_{j_2} + \sum_{j_1 \leqslant j_2 \leqslant j_3} \gamma_{j_1 j_2 j_3} \xi_{j_1} \xi_{j_2} \xi_{j_3} \\ &+ \cdots + \sum_{j_1 \leqslant \cdots \leqslant j_p} \gamma_{j_1 \ldots j_p} \xi_{j_1} \ldots \xi_{j_p}, \end{aligned} \tag{6}$$

where the terms in this representation are self-explanatory.

The interpretation of these polynomial models is complex. The presence of a $j$th-order interaction term means that the joint values of the predictor process at $j$ time-points have an effect on the outcome, in addition to the joint effects of the process values at $\ell$ time-points for all $\ell < j$. For the quadratic model, the interaction effects at two time-points are added to the effects at a single time-point. The interaction effects are perhaps easier to understand in terms of the functional principal components as in version (6), where the interpretation is the same as for the conventional polynomial regression model, which includes all possible interaction terms. The functional principal components themselves are projections of the predictor process in the directions determined by the eigenfunctions and accordingly are interpreted in terms of the shape of their corresponding eigenfunctions, often as contrasts between positively and negatively weighted parts of the predictor process (Castro et al., 1986; Jones & Rice, 1992; Izem & Kingsolver, 2005).

For the models that are expressed in terms of the functional principal components of the forms (5), (6), one can easily introduce variations by omitting some of the interaction terms. For example, a noteworthy variation of the functional quadratic model is

$$E(Y \mid X) = \alpha + \sum_k \beta_k \xi_k + \sum_k \gamma_{kk} \xi_k^2. \tag{7}$$

If expressed in the form of (3), model (7) imposes a restriction on the quadratic parameter function $\gamma(s,t)$, which in this case will be of diagonal form $\gamma(s,t) = \sum_k \gamma_{kk}\phi_k(s)\phi_k(t)$. This version of the functional quadratic regression model does not include interaction terms.

### 2·3. *Explicit representations*

The functional population normal equations provide solutions for functional regression models under certain regularity conditions (He et al., 2000). The functional least-squares deviation, expressed in terms of the parameters $\beta_k$ and $\gamma_{k\ell}$ in representation (4), is given by

$$Q\,\{(\beta_k), (\gamma_{k\ell}), k, \ell$$

fixed or random time-points $T_i = (T_{i1}, \ldots, T_{iN_i})^{\mathrm{T}}$ is subject-specific but nonrandom for dense designs, and is a random variable for sparse designs, assumed to be independently and identically distributed as $N$ and independent of all other random variables. The measurement errors $\varepsilon_{ij}$ are assumed to be independent and identically distributed with $E(\varepsilon_{ij}) = 0$, $E(\varepsilon_{ij}^2) = \sigma^2$, and independent of the functional principal components $\xi_{ik}$ in (2), leading to the representation

$$U_{ij} = X_i(T_{ij}) + \varepsilon_{ij} = \mu_X(T_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(T_{ij}) + \varepsilon_{ij}, \quad T_{ij} \in \mathcal{T}. \tag{10}$$

Estimates $\hat{\mu}_X$, $\hat{G}$, $\hat{\lambda}_k$, $\hat{\phi}_k$ and $\hat{\sigma}^2$ of the underlying population mean function $\mu_X$, covariance function $G$, eigenvalues $\lambda_k$, eigenfunctions $\phi_k$ and error variance $\sigma^2$ are easily obtained by applying a nonparametric functional approach (Yao et al., 2005a), implemented in the PACE package, available at http://www.stat.ucdavis.edu/~mueller/; compare also Rice & Silverman (1991).

A key step is estimation of the regression parameter functions $\beta$ and $\gamma$, based on representations (4) and (9). The cross-covariance surfaces

$$\begin{aligned} C_1(t) &= \operatorname{cov}\{X(t), Y\} = \sum_{k=1}^{\infty} \eta_k \phi_k(t), \quad t \in \mathcal{T}, \\ C_2(s,t) &= E\{X(s)X(t)Y\} = \sum_{k,\ell=1}^{\infty} \rho_{k\ell}\phi_k(s)\phi_\ell(t), \quad s,t \in \mathcal{T}, \end{aligned} \tag{11}$$

can be estimated by using raw covariances $C_i^{(1)}(T_{ij}) = \{U_{ij} - \hat{\mu}_X(T_{ij})\}Y_i$, $1 \leqslant j \leqslant N_i$, and $C_i^{(2)}(T_{ij}, T_{il}) = \{U_{ij} - \hat{\mu}_X(T_{ij})\}\{U_{il} - \hat{\mu}_X(T_{il})\}Y_i$, $1 \leqslant j \neq l \leqslant N_i$, as input for one- and two-dimensional smoothing steps; see (A2) in the Appendix for further details. One needs to remove the diagonal elements $C_i^{(2)}(T_{ij}, T_{ij})$ prior to this smoothing step in order not to contaminate the estimates with the measurement error in $U_{ij}$, in analogy to the situation for autocovariance surface estimation (Yao et al., 2005a). The resulting estimates are denoted by $\hat{C}_1$ and $\hat{C}_2$. The bandwidths for the one- and two-dimensional smoothing steps needed to obtain $\hat{C}_1$ and $\hat{C}_2$ are chosen by generalized crossvalidation, similarly to the choices implemented in PACE; in related studies, the resulting estimation errors were found to be not overly sensitive to these choices; see, e.g. Liu & Müller (2009).

From (11) one then obtains estimates of the quantities $\eta_k$ and $\rho_{k\ell}$ in (9), $k, \ell = 1, \ldots, K$, where $K$ is the number of eigenfunctions included for approximating the predictor process $X$,

$$\hat{\eta}_k = \int_{\mathcal{T}} \hat{C}_1(t)\hat{\phi}_k(t)dt, \quad \hat{\rho}_{k\ell} = \int_{\mathcal{T}}\int_{\mathcal{T}} \hat{C}_2(s,t)\hat{\phi}_k(s)\hat{\phi}_\ell(t)\,dsdt \quad (k, \ell = 1, \ldots, K), \tag{12}$$

by observing that the relations in (12) hold for the corresponding population quantities. Using $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ and plugging in the estimates (12), one then obtains estimates of the regression coefficients in (5):

$$\hat{\alpha} = \bar{Y} - \sum_{k=1}^{K} \hat{\gamma}_{kk}\hat{\lambda}_k, \quad \hat{\beta}_k = \hat{\lambda}_k^{-1}\hat{\eta}_k, \quad \hat{\gamma}_{k\ell} = (\hat{\lambda}_k\hat{\lambda}_\ell)^{-1}\hat{\rho}_{k\ell} \quad (k, \ell = 1, \ldots, K). \tag{13}$$

Regarding estimation of $\gamma_{kk}$, for dense designs, we use the moment estimates $\hat{\tau}_{D,k} = n^{-1}\sum_{i=1}^{n} \hat{\xi}_{I,ik}^4$ for fourth moments $\tau_k$, where $\hat{\xi}_{I,ik}$ is based on the integral method (15).

Neither the proposed estimation schemes nor the consistency results require Gaussianity for the dense design case. The situation is different for the sparse case, where the integration-based estimates $\hat{\xi}^{I}_{I,ik}$, used for the estimation of $\gamma_{kk}$, are not consistent, due to the sparseness of the design. This difficulty can be overcome by making the assumption that $\tau_k = 3\lambda_k^2\,(k = 1, 2, \ldots)$, which then makes it possible to extend the above estimation scheme to the sparse case, yielding

$$\hat{\gamma}_{D,kk} = (\hat{\rho}_{kk} - \bar{Y}\hat{\lambda}_k)/(\hat{\tau}_k - \hat{\lambda}_k^2), \quad \hat{\gamma}_{S,kk} = \hat{\lambda}_k^{-2}(\hat{\rho}_{kk} - \bar{Y}\hat{\lambda}_k)/2, \tag{14}$$

where subscripts $D$ and $S$ denote dense and sparse designs. The resulting estimates for the regression functions are

$$\hat{\beta}(t) = \sum_{k=1}^{K} \hat{\beta}_k \hat{\phi}_k(t), \quad \hat{\gamma}(s,t) = \sum_{k=1}^{K}\sum_{\ell=1}^{k} \hat{\gamma}_{k\ell}\hat{\phi}_k(s)\hat{\phi}_\ell(t), \quad s, t \in \mathcal{T},$$

where $\hat{\beta}_k$ and $\hat{\gamma}_{k\ell}$ are as in (13), and $\hat{\gamma}_{kk}$ is as in (14).

In the sparse case, for the simpler functional linear model (1), Gaussianity or other restrictive assumptions are not needed for consistent estimation of the regression parameter function $\beta$. However, in the quadratic case, sparse designs require the additional assumption $\tau_k = 3\lambda_k^2\,(k = 1, 2, \ldots)$ for consistent estimation of the parameter $\tau_k$; this assumption is satisfied under Gaussianity, which, however, is not required at this stage. For data-based choice of the number of included components $K$, a variety of options is available, including crossvalidation or a variant of the Bayesian information criterion; we adopt the latter; see (A3) in the Appendix.

### 3·2. *Prediction*

We next aim for the prediction of an unknown response $Y^*$, based on noisy observations $U^* = (U_1^*, \ldots, U_{N^*}^*)^{\mathrm{T}}$ of a new predictor trajectory $X^*(\cdot)$, taken at $T^* = (T_1^*, \ldots, T_{N^*}^*)^{\mathrm{T}}$. For the dense design case, the traditional integral estimates of functional principal components $\xi_k^*$, based on the definition $\xi_k^* = \int\{X^*(t) - \mu_X(t)\}\phi_k(t)\,dt$, are

$$\hat{\xi}^*_{I,k} = \sum_{j=2}^{N^*}\{U_j^* - \hat{\mu}_X(T_j^*)\}\hat{\phi}_k(T_j^*)(T_j^* - T_{j-1}^*) \quad (k = 1, \ldots, K). \tag{15}$$

The interaction and quadratic terms, as needed for the functional quadratic model, are obtained directly by $\hat{\xi}^*_{I,k}\hat{\xi}^*_{I,\ell}\,(k, \ell = 1, \ldots, K)$.

Considering the sparse design case, define $\xi^* = (\xi_1^*, \ldots, \xi_K^*)^{\mathrm{T}}$, $\phi_k^* = \{\phi_k(T_1^*), \ldots, \phi_k(T_{N^*}^*)\}^{\mathrm{T}}$, the $K \times N^*$ matrix $H = (\lambda_1\phi_1^*, \ldots, \lambda_K\phi_K^*)^{\mathrm{T}}$ and $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_K\}$. The best prediction of the elements of the matrix $\xi^*\xi^{*\mathrm{T}}$, given $U^*$, is the conditional expectation $E(\xi^*\xi^{*\mathrm{T}} \mid U^*)$, for which estimates under Gaussian assumptions are given by

$$(\widehat{\xi_k^*\xi_\ell^*})_{P,1\leqslant k,\ell\leqslant K} = \hat{E}(\xi^*\xi^{*\mathrm{T}} \mid U^*) = \hat{\xi}^*\hat{\xi}^{*T} + \hat{\Lambda} - \hat{H}\hat{\Sigma}_{U^*}^{-1}\hat{H}^{\mathrm{T}}. \tag{16}$$

Here $\hat{\xi}^* = (\hat{\xi}^*_{P,1}, \ldots, \hat{\xi}^*_{P,K})^{\mathrm{T}}$ is a vector of estimates $\hat{\xi}^*_{P,k}$ as in (A1), obtained in a conditioning step. The estimate $\hat{\Sigma}_{U^*}$ of $\Sigma_{U^*}$ is obtained by substituting estimates $\hat{G}$ and $\hat{\sigma}^2$ for $G$ and $\sigma^2$.

For both dense and sparse designs, the functional quadratic prediction of the response $Y^*$ from the measurements $U^*$ is then given by

$$\hat{Y}^* = \hat{\alpha} + \sum_{k=1}^{K}\hat{\beta}_k\hat{\xi}_k^* + \sum_{k=1}^{K}\sum_{\ell=1}^{k}\hat{\gamma}_{k\ell}\widehat{\xi_k^*\xi_\ell^*}, \tag{17}$$

where $\hat{\xi}_k^*$, $\widehat{\xi_k^*\xi_\ell^*}$ refer to $\hat{\xi}^*_{I,k}$, $\hat{\xi}^*_{I,k}\hat{\xi}^*_{I,\ell}$ as in (15) for dense designs and to $\hat{\xi}^*_{P,k}$ (A1), $(\widehat{\xi_k^*\xi_\ell^*})_P$ (16) for sparse designs, while $\hat{\alpha}$, $\hat{\beta}_k$ and $\hat{\gamma}_{k\ell}$ are obtained as in (13) and (14).

In many applications a simple empirical measure to gauge the strength of the regression relation is useful. One such measure that coincides with the usual coefficient of determination in a simple linear regression, and in general provides a comparison of the prediction error when using a simple sample mean of the responses for prediction with that using a proposed predictor is the following quasi-$R^2$:

$$\hat{R}_Q^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \tag{18}$$

where the predicted response $\hat{Y}_i$ for the $i$th subject is as in (17). This quasi-$R^2$ does not automatically increase when predictors are added to a model and permits straightforward interpretation and model comparison. The estimation scheme we have outlined for functional quadratic regres-

$\pi_k = 1/\lambda_k + 1/\min_{1\leqslant j\leqslant k}(\lambda_j - \lambda_{j+1})$,

$$\hat{\alpha} - \alpha = O_p\left(\frac{1}{n^{1/2}h^2}\sum_{k=1}^{K}\frac{\pi_k}{\lambda_k^2} + \frac{\mathbb{I}_D K}{\bar{N}^{1/2}}\right) + o_p\left\{\left(\sum_{k=K+1}^{\infty}\gamma_{kk}^2\right)^{1/2}\right\}, \tag{20}$$

$$\|\hat{\beta} - \beta\| = O_p\left(\frac{1}{n^{1/2}h^2}\sum_{k=1}^{K}\frac{\pi_k}{\lambda_k} + R_{\beta,K}\right), \tag{21}$$

$$\|\hat{\gamma} - \gamma\|_{\mathcal{T}^2} = O_p\left(\frac{1}{n^{1/2}h^2}\sum_{1\leqslant\ell\leqslant k\leqslant K}\frac{\pi_k + \pi_\ell}{\lambda_k\lambda_\ell} + \frac{\mathbb{I}_D K^2}{\bar{N}^{1/2}} + R_{\gamma,K}\right). \tag{22}$$

We next consider the consistency of the prediction of $Y^*$ for a new subject or sampling unit. For dense designs, the prediction given the data $(U_1^*, \ldots, U_{N^*}^*)$ targets $E(Y^* \mid X^*)$ as in (5). For sparse designs, due to the sparsity of the available measurements $U^*$ for the predictor trajectory $X^*$, the target of the prediction is conditional on these measurements and thus becomes

$$\tilde{Y}^* = E\{E(Y^* \mid X^*) \mid U^*\} = \alpha + \sum_{k=1}^{\infty}\beta_k E(\xi_k^* \mid U^*) + \sum_{1\leqslant\ell\leqslant k\leqslant\infty}\gamma_{k\ell}E(\xi_k^*\xi_\ell^* \mid U^*). \tag{23}$$

THEOREM 2. *Let $\hat{Y}^*$ be the prediction* (17) *for both dense and sparse designs, $E(Y^* \mid X^*)$ as in* (5)*, and $\tilde{Y}^*$ as in* (23).

(i) *Under* (A1)–(A3) *for dense designs, as $n \to \infty$,*

$$\hat{Y}^* - E(Y^* \mid X^*) = O_p\left(\frac{1}{n^{1/2}h^2}\sum_{1\leqslant\ell\leqslant k\leqslant K}\frac{\pi_k + \pi_\ell}{\lambda_k\lambda_\ell} + \frac{K^2}{N^{*1/2}}\right) + o_p(R_{\beta,K} + R_{\gamma,K}). \tag{24}$$

(ii) *Under* (A2) *and* (A4) *for sparse designs, as $n \to \infty$,*

$$\hat{Y}^* - \tilde{Y}^* = O_p\left(\frac{1}{n^{1/2}h^2}\sum_{1\leqslant\ell\leqslant k\leqslant K}\frac{\pi_k + \pi_\ell}{\lambda_k\lambda_\ell}\right) + o_p(R_{\beta,K} + R_{\gamma,K}). \tag{25}$$

This result establishes the consistency of the prediction of the response, given the data for a new subject. Extensions to more general polynomial models are analogous.

## 5. SIMULATION STUDIES

We studied the Monte Carlo performance of the functional quadratic model (3) in comparison with the functional linear model (1) for both dense and sparse designs. Each of the 400 simulation runs consisted of a sample of $n = 100$ predictor trajectories $X_i$, with mean function $\mu_X(s) = s + \sin(s)$ $(0 \leqslant s \leqslant 10)$ and a covariance function derived from two eigenfunctions, $\phi_1(s) = -5^{-1/2}\cos(\pi s/10)$ and $\phi_2(s) = 5^{-1/2}\sin(\pi s/10)$ $(0 \leqslant s \leqslant 10)$. The corresponding eigenvalues were chosen as $\lambda_1 = 4$, $\lambda_2 = 1$ and $\lambda_k = 0$ $(k \geqslant 3)$, the measurement errors in (10) as $\varepsilon_{ij} \sim N(0, 0{\cdot}5^2)$. To study the effect of Gaussianity, which is of interest especially for the sparse design case, we considered two settings: (i) $\xi_{ik} \sim \mathcal{N}(0, \lambda_k)$, a Gaussian case; (ii) $\xi_{ik}$ are generated from the mixture of two normals, $\mathcal{N}\{(\lambda_k/2)^{1/2}, \lambda_k/2)$ with probability 1/2 and

Table 1. *Monte Carlo estimates of the* 25*th,* 50*th and* 75*th percentiles of relative prediction error, comparing predictions obtained by the functional quadratic model and functional linear model for both dense and sparse designs, based on* 400 *Monte Carlo runs with sample size* $n = 100$. *The underlying regression function is quadratic or linear, and the principal components of the predictor process are generated from Gaussian or mixture distributions*

| | | | Gaussian | | | Mixture | | |
|---|---|---|---|---|---|---|---|---|
| Design | True | Fitted | 25th | 50th | 75th | 25th | 50th | 75th |
| | | FQM | 0·037 | 0·201 | 1·403 | 0·033 | 0·166 | 0·950 |

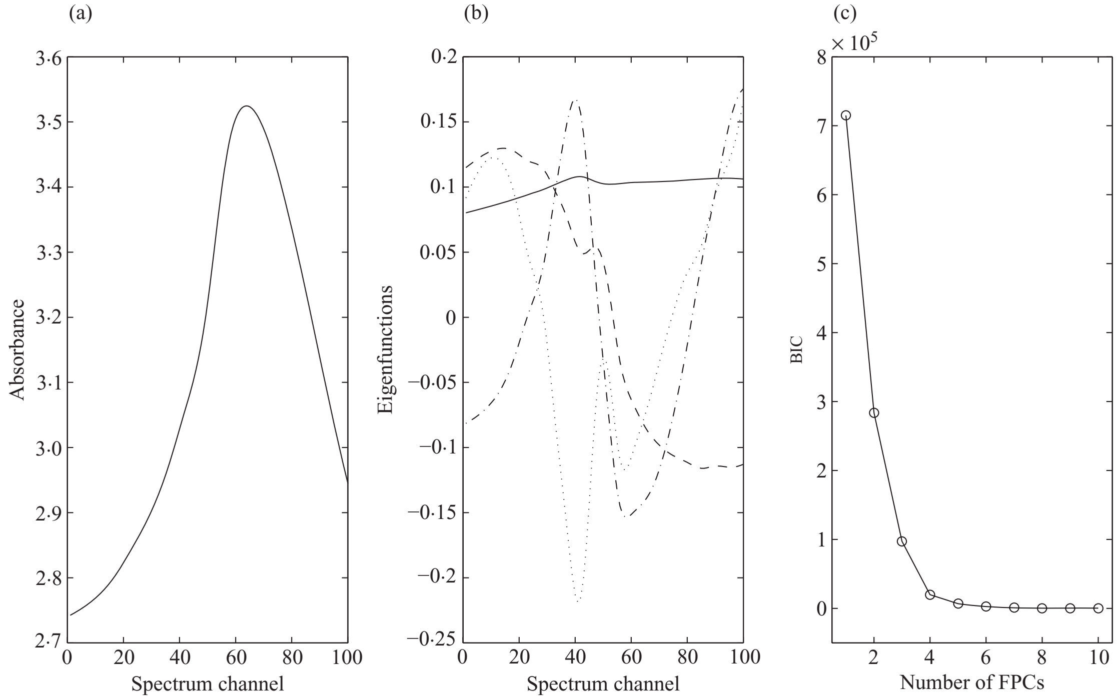


Fig. 2. Smooth estimates of the predictor mean function (a) and first (solid), second (dashed), third (dashed-dot) and fourth (dotted) eigenfunctions (b), as well as the values of the Bayesian information criterion (A3), plotted against the number of included functional principal components (FPCs) (c).

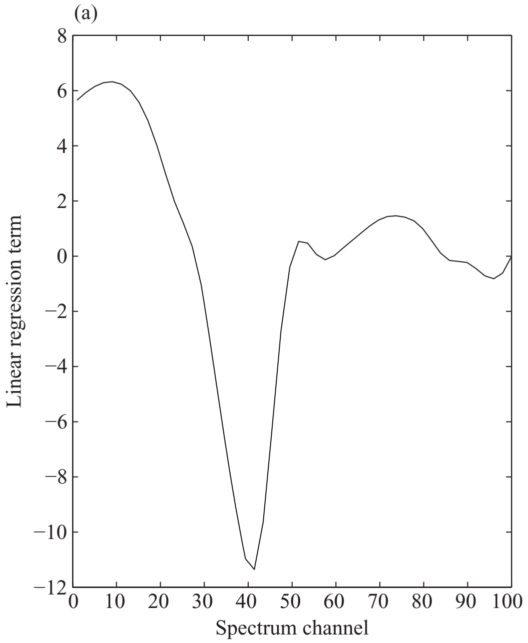
The task is to predict fat contents from the spectrum, setting the stage for a functional regression analysis.

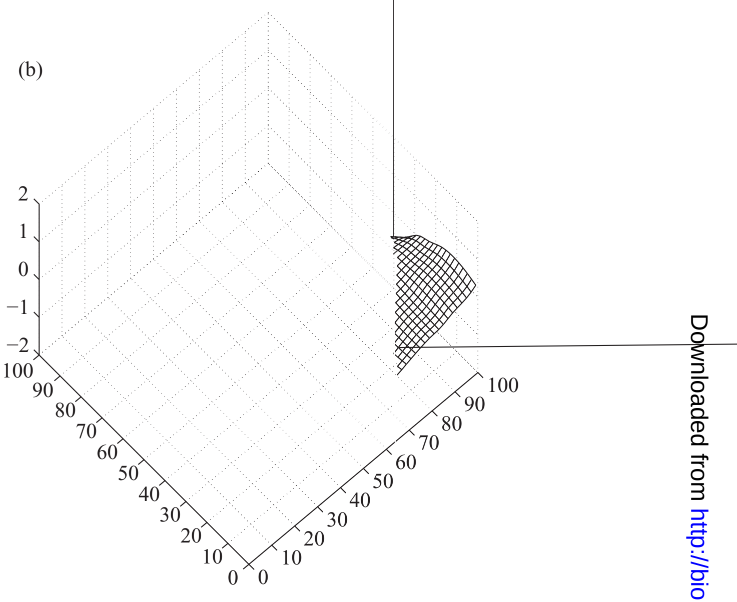
A subsample of 50 randomly selected spectra is displayed in Fig. 1, indicating that these predictor trajectories are smooth. The estimated mean function is in Fig. 2(a). Four eigenfunctions are selected for modelling, explaining more than 99.8% of the total variation, and are visualized in Fig. 2(b). The estimated univariate linear function $\beta(t)$ and the bivariate quadratic surface $\gamma(s, t)$ in Fig. 3 each exhibit several broad peaks and valleys; especially spectral values near 40 units are strongly weighted. The quadratic response surface obtained from the fitted quadratic regression model is presented in Fig. 4.

For prediction, one typically will choose additional components, guided by one-leave-out prediction errors. We compare the prediction performance of the proposed functional quadratic model with functional linear regression and also with partial least-squares, a popular approach in chemometrics; we refer to Xu et al. (2007) and the references therein. The results for prediction errors and quasi-$R^2$ (18) in dependence on the number of included components in Table 2 demonstrate that for more than three components, as required for a reasonably good prediction, the error of the functional quadratic model is consistently smaller than that for partial least-squares, which in turn is smaller than that of functional linear regression.

## Acknowledgement

We wish to thank two referees and the editor for constructive comments. This research was supported by the National Science Foundation, U.S.A., and a Discovery grant from the Natural Sciences and Engineering Council of Canada.





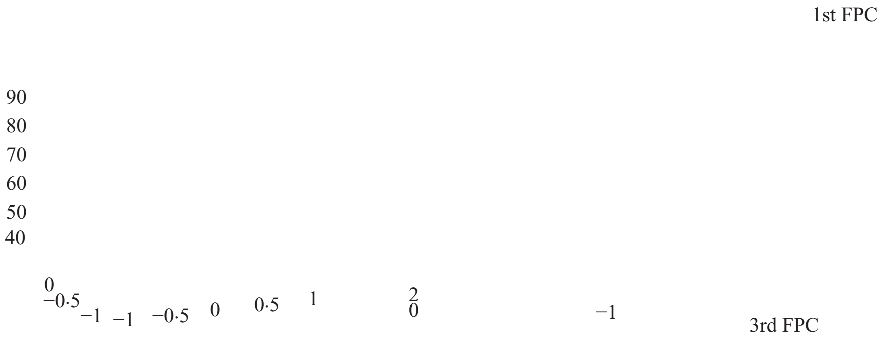


Fig. 4. Fitted quadratic response surface for $\hat{Y}$ (fat content), obtained by varying two predictor functional principal components, FPCs, at a time and fixing the other two at 0. Arranged from top left to right and then bottom left to right: $\hat{Y}$ versus $\xi$ $_1, \xi_2$ 1 3 1 4 $_2, \xi_3,$ $\xi$ 0), (0 2 4 $_3, \xi_4$

Table 2. *Medians of crossvalidated relative prediction errors* ($\times 10^4$) *and of quasi-*$R^2$ (18) ($R_Q^2 \times 100$)*, for varying numbers of components, comparing the functional quadratic model, the functional linear model and partial least-squares for spectral data*

| Components | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FQM | PE | 101 | 84·4 | 16·9 | 6·58 | 2·30 | 1·79 | 1·08 | 1·18 | 0·60 | 0·50 |
| | $R_Q^2$ | 24·6 | 32·5 | 77·8 | 94·1 | 98·5 | 97·8 | 99·0 | 99·3 | 99·6 | 99·8 |
| FLM | PE | 81·3 | 64·7 | 35·0 | 11·9 | 5·31 | 5·72 | 5·05 | 5·57 | 4·75 | 5·58 |
| | $R_Q^2$ | 22·3 | 27·9 | 70·2 | 90·1 | 93·8 | 94·1 | 94·4 | 94·4 | 94·5 | 94·5 |
| PLS | PE | 80·1 | 26·0 | 12·5 | 10·1 | 6·40 | 5·87 | 5·68 | 5·05 | 4·10 | 4·04 |
| | $R_Q^2$ | 23·1 | 76·5 | 86·9 | 90·8 | 93·8 | 94·2 | 94·6 | 94·8 | 96·4 | 96·5 |

FQM, functional quadratic model; FLM, functional linear model; PLS, partial least-squares.

## Appendix

### A1. *Estimation procedures*

Using the notation introduced in § 3, the estimates of $\xi_k^*$ obtained by the principal analysis by conditional expectation procedure (Yao et al., 2005a) are given by

$$\hat{\xi}_{P,k}^* = \hat{\lambda}_k \hat{\phi}_k^{*\mathrm{T}} \hat{\Sigma}_{U^*}^{-1}(U^* - \mu_X^*) \quad (k = 1, \ldots, K), \tag{A1}$$

where $\mu_X^* = (\mu_X(T_1^*), \ldots, \mu_X(T_m^*))^{\mathrm{T}}$. It follows from results in Müller (2005) that, as designs become dense, $\hat{\xi}_{I,k}^*$ (15) and $\hat{\xi}_{P,k}^*$ (A1) can be considered asymptotically equivalent. Smoothing kernels $\kappa_1, \kappa_2$ are compactly supported smooth densities with zero means and finite variances and are implemented with suitable bandwidth sequences $b$ and $h$. With $f\{\theta, (s, t), (T_{ij}, T_{il})\} = \theta_0 + \theta_{11}(s - T_{ij}) + \theta_{12}(t - T_{il})$, the one- and two-dimensional smoothers to estimate $C_1(t)$ and $C_2(s, t)$ in (11) are obtained by minimizing

$$\sum_{i=1}^{n} \sum_{j=1}^{N_i} \kappa_1\left(\frac{T_{ij} - t}{b}\right) \left\{C_i^{(1)}(T_{ij}) - \alpha_0 - \alpha_1(t - T_{ij})\right\}^2,$$

$$\sum_{i=1}^{n} \sum_{1 \leqslant j \neq l \leqslant N_i} \kappa_2\left(\frac{T_{ij} - s}{h}, \frac{T_{il} - t}{h}\right) \left[C_i^{(2)}(T_{ij}, T_{il}) - f\{\theta, (s, t), (T_{ij}, T_{il})\}\right]^2, \tag{A2}$$

with respect to $\alpha = (\alpha_0, \alpha_1)^{\mathrm{T}}$ and $\theta = (\theta_0, \theta_{11}, \theta_{12})^{\mathrm{T}}$, yielding $\hat{C}_1(t) = \hat{\alpha}_0(t)$ and $\hat{C}_2(s, t) = \hat{\theta}_0(s, t)$. The number of included components is chosen by minimizing

$$\mathrm{BIC}(K) \propto \sum_{i=1}^{n} \sum_{j=1}^{N_i} \left[-\frac{1}{2\hat{\sigma}^2} \left\{U_{ij} - \hat{\mu}(T_{ij}) - \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\phi}_k(T_{ij})\right\}^2\right] + K \log\left(\sum_{i=1}^{n} N_i\right). \tag{A3}$$

### A2. *Technical assumptions and proofs*

For model (3) or (5) to be well defined in the least-squares sense, we require the following moment conditions for predictor processes. Let $\nu_1$ and $\nu_2$ be positive integers.

(A1) Assume that $\sum_{k=1}^{\infty} E\xi_k^4 < \infty$, $E(\xi_k^{\nu_1} \xi_\ell^{\nu_2}) = E\xi_k^{\nu_1} E\xi_\ell^{\nu_2}$ for $\nu_1 + \nu_2 = 3$ and $1 \leqslant k, \ell < \infty$; $E(\xi_k^{\nu_1} \xi_\ell^{\nu_2}) = E\xi_k^{\nu_1} E\xi_\ell^{\nu_2}$ for $\nu_1 + \nu_2 = 4$ and $1 \leqslant k \neq \ell < \infty$.

Let $b^* = b^*(n)$, $h^* = h^*(n)$, $\tilde{h} = \tilde{h}(n)$ denote the bandwidths for estimating $\hat{\mu}_X$ (24), $\hat{G}$ (25) and $\hat{\sigma}$ (2) in Yao et al. (2005a). The Fourier transforms of $\kappa_1$ and $\kappa_2$ are given by $\kappa_{F,1}(u) = \int \exp(-iut)\kappa_1(t)\,dt$ and $\kappa_{F,2}(u, v) = \int \exp\{-(iut + ivs)\}\kappa_2(s, t)\,ds\,dt$, respectively.

(A2·1) Assume that $\max(b^*, h^*, \tilde{h}, b, h) \to 0$, $\min(nb^{*4}, n\tilde{h}^4, nb^4) \to \infty$, $\max(nb^{*6}, n\tilde{h}^6, nb^6) < \infty$, $\min(nh^{*6}, nh^6) \to \infty$, $\max(nh^{*8}, nh^8) < \infty$, as $n \to \infty$ and $h = O\{\min(b^{1/2}, b_1^{*,1/2}, \tilde{h}^{1/2}, h^*)\}$.

(A2·2) Assume that $\kappa_{F,1}$ and $\kappa_{F,2}$ are absolutely integrable, $\int|\kappa_{F,1}(u)|du<\infty$, $\int\int|\kappa_{F,2}(u,v)|dudv<\infty$.

(A2·3) Assume that $n^{-1/2}h^{-2}\sum_{k=1}^{\tilde{K}}\sum_{\ell=1}^{k}(\pi_k+\pi_\ell)(\lambda_k\lambda_\ell)^{-1}\to 0$, as $n\to\infty$, where

$$\begin{aligned} D_X&=\int_{\mathcal{T}^2}\{\hat{G}(s,t)-G(s,t)\}^2dsdt, \qquad \delta_k=\min_{1\leqslant j\leqslant k}(\lambda_j-\lambda_{j+1}),\\ \tilde{K}&=\inf\{j\geqslant 1:\lambda_j-\lambda_{j+1}\leqslant 2D_X\}-1,\ \pi_k=1/\lambda_k+1/\delta_k. \end{aligned} \tag{A4}$$

To obtain consistent functional principal component estimates for dense designs, we require both the pooled data across all subjects and the data from each subject to be dense in $\mathcal{T}$. Denote the sorted time-points across all subjects by $a_0\leqslant T_{(1)}\leqslant\cdots\leqslant T_{(\tilde{N})}\leqslant b_0$, and let $\Delta=\max\{T_{(m)}-T_{(m-1)}:m=1,\ldots,\tilde{N}+1\}$, where $\tilde{N}=\sum_{i=1}^{n}N_i$, $\mathcal{T}=[a_0,b_0]$, $t_{(0)}=a_0$ and $t_{(N+1)}=b_0$. For the $i$th subject, suppose that the time-points $T_{ij}$ have been ordered nondecreasingly. Let $\Delta_i=\max\{T_{ij}-T_{i,j-1}:j=1,\ldots,N_i+1\}$ and $\Delta^*=\max\{\Delta_i:i=1,\ldots,n\}$, where $t_{i0}=a_0$ and $t_{i,n_i+1}=b_0$, and $\bar{N}=\tilde{N}/n$. Put $N_{\max}=\max\{N_i:i=1,\ldots,n\}$ and $N_{\min}=\min\{N_i:i=1,\ldots,n\}$. Denote the distribution that generates $U_{ij}$ for the $i$th subject at $T_{ij}$ by $U_i(t)\sim U(t)$ with density $g_U(u;t)$. Let $g_U^*(u_1,u_2;\,t_1,t_2)$ be the density of $(U(s_1),U(t_2))$ and $\|f\|_\infty=\sup_{t\in\mathcal{T}}|f(t)|$ for any function with support $\mathcal{T}$. The next assumptions are for the case of dense designs, where (A3.3) is needed for consistent estimation of $\tau=E(\xi_k^4)$ and (A3.4) for consistency of the prediction.

(A3·1) Assume that $\Delta=O\{\min(n^{-1/2}b^{*-1},n^{-1/2}\tilde{h}^{-1},n^{-1/2}b^{-1},n^{-1/4}h^{*-1},n^{-1/4}h^{-1})\}$, $\Delta^*=O(1/\bar{N})$, $C_1\bar{N}\leqslant N_{\min}\leqslant N_{\max}\leqslant C_2\bar{N}$ for some $C_1,C_2>0$ and $\sup_{t\in\mathcal{T}}E\{U^4(t)\}<\infty$.

(A3·2) Assume that $(d^2/dt^2)g_U(u;t)$ is uniformly continuous on $\Re\times\mathcal{T}$ and that $\{d^2/(dt_1^{\ell_1}dt_2^{\ell_2})\}$ $g_U^*(u_1,u_2;t_1,t_2)$ is uniformly continuous on $\Re^2\times\mathcal{T}^2$, for $\ell_1+\ell_2=2,\ 0\leqslant\ell_1,\ell_2\leqslant 2$.

(A3·3) Assume that $\tilde{K}^2=o(\bar{N}^{1/2})$, $\max_{k\leqslant\tilde{K}}\|\phi_k'\|_\infty=O(\bar{N}^{1/2})$, $E(\|X'\|_\infty)<\infty$ and $E(\|X'^2\|_\infty^2)=o(\bar{N})$, where $\tilde{K}$ is as in (A4).

(A3·4) Assume that $\tilde{K}^2=o(N^{*1/2})$ and $\max_{k\leqslant\tilde{K}}\|\phi_k'\|_\infty=O(N^{*1/2})$.

For sparse designs, denote the marginal and joint densities of $T$, $(T,U)$ and $(T_1,T_2,U_1,U_2)$ by $g_T(t)$, $g_U(t;u)$, $g_U^*(t_1,t_2;\,u_1,u_2)$. The following assumptions are only needed for sparse designs; (A4·1) and (A4·2) guarantee basic regularity and smoothness requirements, while the Gaussian assumption (A4·3) is needed for consistency of predictions

n_2.2526 159d626 1 Tf 6.9738 0 0 62o1.0 .838 2776.762 Tm (2292d-0.00292dTd dom00293(v)26_Tmiabgula

*and as a consequence,* $\hat\sigma^2 - \sigma^2 = O_p(n^{-1/2}h^{*-2} + n^{-1/2}\tilde h^{-1})$. *Considering eigenvalues* $\lambda_k$ *of multiplicity one,* $\hat\phi_k$ *can be chosen such that*

$$pr\left(\sup_{1\leqslant k\leqslant \tilde K} |\hat\lambda_k - \lambda_k| \leqslant D_X\right) = 1, \quad \sup_{t\in\mathcal{T}} |\hat\phi_k(t) - \phi_k(t)| = O_p\left(\frac{\pi_k}{n^{1/2}h^{*2}}\right) \quad (k = 1, \ldots, \tilde K).$$

Lemma A2. *Under* (A1), (A2), (A3·1)–(A3·3) *for dense designs or under* (A1), (A2), (A4·1) *and* (A4·2) *for sparse designs, it holds for any* $K \leqslant \tilde K$ *that*

$$\hat\beta_k - \beta_k = O_p\left(\frac{\pi_k}{n^{1/2}h^2\lambda_k}\right), \quad \hat\gamma_{k\ell} - \gamma_{k\ell} = O_p\left(\frac{\pi_k + \pi_\ell}{n^{1/2}h^2\lambda_k\lambda_\ell}\right) \quad (k, \ell = 1, \ldots, K), \tag{A5}$$

$$\hat\gamma_{D,kk} - \gamma_{kk} = O_p\left(\frac{\pi_k}{n^{1/2}h^2\lambda_k^2} + \frac{1}{\bar N^{1/2}}\right), \quad \hat\gamma_{S,kk} - \gamma_{kk} = O_p\left(\frac{\pi_k}{n^{1/2}h^2\lambda_k^2}\right) \quad (k = 1, \ldots, K), \tag{A6}$$

*where* $\hat\beta_k$, $\hat\gamma_{k\ell}$ *are as in* (13) *for* $\ell < k$, *and* $\hat\gamma_{D,kk}$, $\hat\gamma_{S,kk}$ *are as in* (14).

*Proof of Lemma* A2. It is easy to show the rate for $\hat\beta_k$, by observing that $\lambda_k^{-1} < \pi_k$ from (A4) and $\hat\eta_k - \eta_k = O_p(\pi_k n^{-1/2}h^{-2})$, $\hat\lambda_k^{-1} - \lambda_k^{-1} = O_p(n^{-1/2}h^{*-2}\lambda_k^{-2})$ from Lemma A1. Regarding $\hat\gamma_{k\ell}$ for $k > \ell$, note that $\hat\rho_{k\ell} - \rho_{k\ell} = O_p\{(\phi_k + \pi_\ell)n^{-1/2}h^{-2}\}$ and $(\hat\lambda_k\hat\lambda_\ell)^{-1} - (\lambda_k\lambda_\ell) = O_p\{n^{-1/2}h^{*-2}$