Submitted to Biometrika (2009), **xx**, x, *pp.* 1–21

Printed in Great Britain

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

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 then based on this equivalence and on basis representations of predictor processes. We also obtain an explicit representation of the re-

* Corresponding author

gression surface that defines quadratic functional regression and provide functional asymptotic results for an increasing number of model components as sample size (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 which includes absorption spectra as functional predictors.

Some key words: Absorption Spectra; Asymptotics; Functional Data Analysis; Polynomial Regression; Prediction; Principal Component.

1. INTRODUCTION

Data which 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 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 Figure 1. Displayed are 50 randomly selected 100-channel absorption spectra which are used to predict the composition of food samples. The spectra shown are from meat specimen 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, with only small measurement errors. Taking advantage of the smoothness of these trajectories is key to the efficient modeling of regression relationships that include functional predictors. The functional nature of the predictors and the measurements need to be adequately reflected in the statistical modeling 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., 2005c). 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 vector data, where quadratic models are commonplace and occupy a well-established place in many applications such as response-surface analysis (see e.g. Melas et al., 2003).

Within the field of functional data analysis, with excellent overviews provided in Ramsay & Silverman (2002, 2005), an emphasis has been the functional linear model. Various aspects of this model have been studied, including implementations, computing and asymptotic theory (Cardot et al., 2003; Shen & Faraway, 2004; Cardot & Sarda, 2005; Cai & Hall, 2006; Hall & Horowitz, 2007). The functional linear model has been brought to the fore in Ramsay & Dalzell (1991). It imposes a structural linear constraint on the regression relationship which may or may not be satisfied. The linear constraint is helpful for addressing the problem that the elements of a finite sample of random functions are very far away from each other in the infinite-dimensional function space, but it may be too restrictive for some applications. It is therefore of interest to consider extensions of the functional linear model that convey a reasonable structural constraint, but at the same time enhance its flexibility. We propose here 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 $\mathrm{cov}\{X(s), X(t)\} = G(s,t)$ 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 $\lambda_1 \geq \lambda_2 \geq \ldots$, 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

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

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

To estimate the function β, 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 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 ξ_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 $\mathrm{var}(\xi_k) = \lambda_k$.

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

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

where α is an intercept. The linear part is seen to be the same as 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 $\{\phi_k\}_{k=1,2,\dots}$ of the process X form a complete basis, the regression parameter functions in (3) can be represented in this basis,

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

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), (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 ξ_k of predictor processes X,

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

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 α 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 ($p \geq 3$) functional polynomial model in analogy to (3) as follows,

$$\begin{aligned} E(Y|X) = \alpha &+ \int_{\mathcal{T}} \beta(t) X^c(t) dt + \int_{\mathcal{T}^2} \gamma(s,t) X^c(s) X^c(t) ds dt \\ &+ \int_{\mathcal{T}^3} \gamma_3(t_1, t_2, t_3) X^c(t_1) X^c(t_2) X^c(t_3) dt_1 dt_2 dt_3 + \ldots \\ &+ \int_{\mathcal{T}^p} \gamma_p(t_1, \ldots, t_p) X^c(t_1) \ldots X^c(t_p) dt_1 \ldots dt_p, \end{aligned} \tag{6}$$

where again α is the intercept, and β, γ, γ_j, $3 \leq j \leq p$, are the linear, quadratic, and jth 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,

$$\begin{aligned} E(Y|X) = \alpha &+ \sum_{j_1 \geq 1} \beta_{j_1} \xi_{j_1} + \sum_{j_1 \leq j_2} \gamma_{j_1 j_2} \xi_{j_1} \xi_{j_2} + \sum_{j_1 \leq j_2 \leq j_3} \gamma_{j_1 j_2 j_3} \xi_{j_1} \xi_{j_2} \xi_{j_3} \\ &+ \ldots + \sum_{j_1 \leq \ldots \leq j_p} \gamma_{j_1 \ldots j_p} \xi_{j_1} \ldots \xi_{j_p}, \end{aligned} \tag{7}$$

where the terms in this representation are self-explanatory.

The interpretation of these polynomial models is complex. The presence of a jth order interaction terms 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 ℓ 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 (7), 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), (7), one can easily introduce variations by omitting some of the interaction terms. For example, a noteworthy variation of the functional quadratic model is

$$E(Y|X) = \alpha + \sum_k \beta_k \xi_k + \sum_k \gamma_{kk} \xi_k^2. \tag{8}$$

If expressed in the form of (3), model (8) 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 β_k and $\gamma_{k\ell}$ in representation (4) is given by

$$Q\left\{(\beta_k), (\gamma_{k\ell}), k, \ell = 1, 2, \ldots, k \le \ell\right\} = E\left(Y - \alpha - \sum_k \beta_k \xi_k - \sum_{k \le \ell} \gamma_{kl} \xi_k \xi_\ell\right)^2. \tag{9}$$

One then obtains the normal equations by differentiating Q with respect to the sets of parameters. Using condition (A1) in the Appendix, the functional normal equations lead to the solutions

$$\begin{aligned} \alpha &= \mu_Y - \sum_k \gamma_{kk} \lambda_k, \quad \beta_k = \eta_k / \lambda_k, && \text{for} \quad \eta_k = E(\xi_k Y), \\ \gamma_{k\ell} &= \rho_{k\ell} / (\lambda_k \lambda_\ell), && \text{for} \quad k < \ell, \quad \rho_{k\ell} = E(\xi_k \xi_\ell Y), \\ \gamma_{kk} &= (\rho_{kk} - \mu_Y \lambda_k) / (\tau_k - \lambda_k^2), && \text{for} \quad \tau_k = E(\xi_k^4). \end{aligned} \tag{10}$$

We note that the representation of the linear parameter function $\beta(t)$ that results from (10) is identical to that in the functional linear model, which is $\beta(t) = \sum_{k=1}^{\infty} \eta_k \lambda_k^{-1} \phi_k(t)$.

3. Inference for Functional Quadratic Regression

3·1. *Estimation*

Our starting point for modeling are the actual observations, which consist either of densely spaced and non-random (dense design) or alternatively of sparse and randomly (irregularly) spaced repeated measurements (sparse design) of the predictor trajectories X_i, contaminated with additional measurement errors. The dimension of the vector of noisy recordings $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{iN_i})^T$ of the random trajectory X_i made at (fixed or random time) points $\boldsymbol{T}_i = (T_{i1}, \ldots, T_{iN_i})^T$ is subject-specific but non-random 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 ε_{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 ξ_{ik} in (2), leading to the representation

$$U_{ij} = X_i(T_{ij}) + \varepsilon_{ij} = \mu_X(T_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(T_{ij}) + \varepsilon_{ij}, \qquad T_{ij} \in \mathcal{T}. \tag{11}$$

Estimates $\hat{\mu}_X, \hat{G}, \hat{\lambda}_k, \hat{\phi}_k$ and $\hat{\sigma}^2$ of the underlying population mean function μ_X, covariance function G, eigenvalues λ_k, eigenfunctions ϕ_k and error variance σ^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 β and γ, based on representations (4) and (10). The cross-covariance surfaces

$$C_1(t) = \mathrm{cov}\{X(t), Y\} = \sum_{k=1}^{\infty} \eta_k \phi_k(t), \qquad t \in \mathcal{T},$$

$$C_2(s,t) = E\{X(s)X(t)Y\} = \sum_{k,\ell=1}^{\infty} \rho_{k\ell}\phi_k(s)\phi_k(t), \qquad s,t \in \mathcal{T}, \tag{12}$$

can be estimated by using "raw" covariances $C_i^{(1)}(T_{ij}) = \{U_{ij} - \hat{\mu}_X(T_{ij})\}Y_i,\ 1 \le j \le N_i$, and $C_i^{(2)}(T_{ij}, T_{il}) = \{U_{ij} - \hat{\mu}_X(T_{ij})\}\{U_{il} - \hat{\mu}_X(T_{il})\}Y_i, 1 \le j \ne l \le N_i$, as input for one- and two-dimensional smoothing steps; see (29) in the Appendix for further details. Note that 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 auto-covariance surface estimation (Yao et al., 2005a). The resulting estimates are denoted by $\widehat{C}_1$ and $\widehat{C}_2$.The bandwidths for the one- and two-dimensional smoothing steps needed to obtain $\hat{C}_1$ and $\hat{C}_2$ are chosen by generalized cross-validation, 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 (12) one then obtains estimates of the quantities η_k and $\rho_{k\ell}$ in (10), $k, \ell = 1, \ldots, K$, where K is the number of eigenfunctions included for approximating the predictor process X,

$$\hat{\eta}_k = \int_{\mathcal{T}} \widehat{C}_1(t)\hat{\phi}_k(t)dt, \qquad \hat{\rho}_{k\ell} = \int_{\mathcal{T}} \int_{\mathcal{T}} \widehat{C}_2(s,t)\hat{\phi}_k(s)\hat{\phi}_\ell(t)dsdt, \tag{13}$$

by observing that the relations in (13) hold for the corresponding population quantities. Using $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ and plugging in the estimates (13), one then obtains estimates of the regression coefficients in (5),

$$\hat{\alpha} = \bar{Y} - \sum_{k=1}^{K} \hat{\gamma}_{kk}\hat{\lambda}_k, \;\; \hat{\beta}_k = \hat{\lambda}_k^{-1}\hat{\eta}_k, \;\; \hat{\gamma}_{k\ell} = (\hat{\lambda}_k\hat{\lambda}_\ell)^{-1}\hat{\rho}_{k\ell}, \quad 1 \le \ell < k \le K. \tag{14}$$

Regarding estimation of γ_{kk}, for dense designs, we use the moment estimates $\hat{\tau}_k^D = n^{-1}\sum_{i=1}^n \hat{\xi}_{ik}^{I\,4}$ for fourth moments τ_k, where $\hat{\xi}_{ik}^I$ is based on the integral method (17).

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 $\hat{\xi}_{ik}^I$, used for the estimation of γ_{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

$$\hat{\gamma}_{kk}^D = (\hat{\rho}_{kk} - \bar{Y}\hat{\lambda}_k)/(\hat{\tau}_k - \hat{\lambda}_k^2), \quad \hat{\gamma}_{kk}^S = \hat{\lambda}_k^{-2}(\hat{\rho}_{kk} - \bar{Y}\hat{\lambda}_k)/2, \tag{15}$$

where superscripts ^{D} and ^{S} denote dense and sparse designs. The resulting estimates for the regression functions are

$$\hat{\beta}(t) = \sum_{k=1}^{K} \hat{\beta}_k \hat{\phi}_k(t) \quad \text{and} \quad \hat{\gamma}(s,t) = \sum_{k=1}^{K} \sum_{\ell=1}^{k} \hat{\gamma}_{k\ell} \hat{\phi}_k(s) \hat{\phi}_\ell(t), \quad s,t \in \mathcal{T}, \tag{16}$$

where $\hat{\beta}_k$ and $\hat{\gamma}_{k\ell}$ are as in (14), and $\hat{\gamma}_{kk}$ as in (15).

In the sparse case, for the simpler functional linear model (e.g. Yao et al., 2005b), Gaussianity or other restrictive assumptions are not needed for consistent estimation of the regression parameter function (corresponding to β with the component γ absent). 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 τ_k; this assumption is satisfied under Gaussianity (which is not required at this stage). For the choice of the number of included components K one may use cross-validation or model selection criteria such as pseudo-BIC (Bayesian information criterion); we adopt the latter, see (30) in the appendix.

3·2. *Prediction*

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

$$\hat{\xi}_k^{*I} = \sum_{j=2}^{N^*} \{U_j^* - \hat{\mu}_X(T_j^*)\} \hat{\phi}_k(T_j^*)(T_j^* - T_{j-1}^*), \qquad k = 1, \ldots, K. \tag{17}$$

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

Considering the sparse design case, define $\boldsymbol{\xi}_K^* = (\xi_1^*, \ldots, \xi_K^*)^T$, $\boldsymbol{\phi}_k^* = \{\phi_k(T_1^*), \ldots, \phi_k(T_{N^*}^*)\}^T$, the $K \times N^*$ matrix $H = (\lambda_1 \boldsymbol{\phi}_1^*, \ldots, \lambda_K \boldsymbol{\phi}_K^*)^T$ and $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_K\}$. The best prediction of the elements of the matrix $\boldsymbol{\xi}_K^* \boldsymbol{\xi}_K^{*T}$, given $\boldsymbol{U}^*$, is the conditional expectation $E(\boldsymbol{\xi}_K^* \boldsymbol{\xi}_K^{*T} | \boldsymbol{U}^*)$, which under Gaussian assumptions is found to be

$$(\widehat{\xi_k^* \xi_\ell^*}^P)_{1 \le k, \ell \le K} = \widehat{E}(\boldsymbol{\xi}_K^* \boldsymbol{\xi}_K^{*T} | \boldsymbol{U}^*) = \hat{\boldsymbol{\xi}}_K^* \hat{\boldsymbol{\xi}}_K^{*T} + \widehat{\Lambda} - \widehat{H} \widehat{\Sigma}_{U^*}^{-1} \widehat{H}^T. \tag{18}$$

Here $\hat{\boldsymbol{\xi}}_K^* = (\hat{\xi}_1^{*P}, \ldots, \hat{\xi}_K^{*P})^T$ is a vector of estimates $\hat{\xi}_k^{*P}$ as in (28), obtained in a conditioning step. The estimate $\widehat{\Sigma}_{U^*}$ of Σ_{U^*} is obtained by substituting estimates $\hat{G}$ and $\hat{\sigma}^2$ for G and σ^2.

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

$$\widehat{Y}^* = \hat{\alpha} + \sum_{k=1}^{K} \hat{\beta}_k \hat{\xi}_k^* + \sum_{k=1}^{K} \sum_{\ell=1}^{k} \hat{\gamma}_{k\ell} \widehat{\xi_k^* \xi_\ell^*}, \tag{19}$$

where $\hat{\xi}_k^*$, $\widehat{\xi_k^*\xi_\ell^*}$ refer to $\hat{\xi}_k^{*I}$, $\hat{\xi}_k^{*I}\hat{\xi}_\ell^{*I}$ as in (17) for dense designs and to $\hat{\xi}_k^{*P}$ (28), $\widehat{\xi_k^*\xi_\ell^*}^P$ (18) for sparse designs and $\hat{\alpha}$, $\hat{\beta}_k$ and $\hat{\gamma}_{k\ell}$ are obtained as in (14) and (15).

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

$$\widehat{R}_Q^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \tag{20}$$

where the predicted response $\widehat{Y}_i$ for the ith subject is as in (19). This quasi-R^2 does not automatically increase when predictors are added to a model and permits straightforward interpretation and model comparison. We note that the estimation scheme we have outlined for functional quadratic regression can be analogously extended to the case of functional polynomial models. Such an extension is particularly straightforward for polynomial models that do not include interaction terms such as the variant of the quadratic model in (8).

4. Theoretical Properties

We consider the asymptotic consistency of the estimated regression functions in a functional setting where the number of functional principal components depends on the sample size n, i.e., $K = K(n)$, and tends to infinity as the sample size increases, $n \to \infty$. This allows for increasingly better approximations of the underlying functions for larger samples, adequately reflecting the nonparametric nature of the problem. In practice the choice of K will depend on the intrinsic structural complexity and also estimation accuracy of the covariance structure. Writing $\|\beta\| = \{\int_{\mathcal{T}} \beta^2(t)dt\}^{1/2}$, $\|\gamma\|_{\mathcal{T}^2} = \{\int_{\mathcal{T}}\int_{\mathcal{T}} \gamma^2(s,t)dsdt\}^{1/2}$, $\beta_K(t) = \sum_{k=1}^K \beta_k\phi_k(t)$ and $\gamma_K(s,t) = \sum_{k=1}^K \sum_{\ell=1}^k \gamma_{k\ell}\phi_k(s)\phi_\ell(t)$, we note that the square integrability properties imply, as $n \to \infty$,

$$R_{\beta,K} \equiv \|\beta - \beta_K\| = (\sum_{k=K+1}^{\infty} \beta_k^2)^{1/2} \to 0,$$

$$R_{\gamma,K} \equiv \|\gamma - \gamma_K\|_{\mathcal{T}^2} = (\sum_{k=K+1}^{\infty}\sum_{\ell=1}^{k} \gamma_{k\ell}^2)^{1/2} \to 0. \tag{21}$$

In the following, h denotes the bandwidth for estimating the surface C_2 in (29). The technical assumptions are listed in Appendix A2, where also the sequence $\widetilde{K} = \widetilde{K}(n)$ and the constants π_k, which play a role in the following results, are defined in equation (31). The increasing sequence $\widetilde{K}(n)$ places an upper bound on the number K of predictor components that are included in the functional quadratic regression. The following result provides consistency of the estimated model components defined at (14). For conciseness, let $\mathbb{I}_D$ denote an indicator for terms that are to be included for dense designs, i.e. $\mathbb{I}_D = 1$ for dense and $= 0$ for sparse designs.

THEOREM 1. *Under (A1), (A2), (A3.1)-(A3.3) (see Appendix for all assumptions) for dense designs or under (A1), (A2), (A4.1) and (A4.2) for sparse designs, as* $n \to \infty$*, for any sequence* $K = K(n) \le \widetilde{K}(n)$*, with* $\pi_k = 1/\lambda_k + 1/\min_{1\le j\le k}(\lambda_j - \lambda_{j+1})$,

$$\hat{\alpha} - \alpha = O_p\Big(\frac{1}{n^{1/2}h^2}\sum_{k=1}^{K}\frac{\pi_k}{\lambda_k^2} + \frac{\mathbb{I}_D K}{\bar{N}^{1/2}}\Big) + o_p\Big\{(\sum_{k=K+1}^{\infty}\gamma_{kk}^2)^{1/2}\Big\} \tag{22}$$

$$\|\hat{\beta} - \beta\| = O_p\Big(\frac{1}{n^{1/2}h^2}\sum_{k=1}^{K}\frac{\pi_k}{\lambda_k} + R_{\beta,K}\Big), \tag{23}$$

$$\|\hat{\gamma} - \gamma\|_{\mathcal{T}^2} = O_p\Big(\frac{1}{n^{1/2}h^2}\sum_{1\le\ell\le k\le K}\frac{\pi_k + \pi_\ell}{\lambda_k\lambda_\ell} + \frac{\mathbb{I}_D K^2}{\bar{N}^{1/2}} + R_{\gamma,K}\Big). \tag{24}$$

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 $E(Y^*|X^*)$ as in (5). For sparse designs, due to the sparsity of the available measurements $\boldsymbol{U}^*$ for the predictor trajectory X^*, the target of the prediction is conditional on these measurements and thus becomes

$$\widetilde{Y}^* = E\{E(Y^*|X^*)|\boldsymbol{U}^*\} = \alpha + \sum_{k=1}^{\infty}\beta_k E(\xi_k^*|\boldsymbol{U}^*) + \sum_{1\le\ell\le k\le\infty}\gamma_{k\ell}E(\xi_k^*\xi_\ell^*|\boldsymbol{U}^*). \tag{25}$$

THEOREM 2. *Let $\widehat{Y}^*$ be the prediction (19) for both dense and sparse designs, $E(Y^*|X^*)$ as in (5), and $\widetilde{Y}^*$ as in (25).*

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

$$\widehat{Y}^* - E(Y^*|X^*) = O_p\Big(\frac{1}{n^{1/2}h^2}\sum_{1\le\ell\le k\le K}\frac{\pi_k+\pi_\ell}{\lambda_k\lambda_\ell}+\frac{K^2}{N^{*1/2}}\Big)+o_p(R_{\beta,K}+R_{\gamma,K}). \quad (26)$$

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

$$\widehat{Y}^* - \widetilde{Y}^* = O_p\Big(\frac{1}{n^{1/2}h^2}\sum_{1\le\ell\le k\le K}\frac{\pi_k+\pi_\ell}{\lambda_k\lambda_\ell}\Big)+o_p(R_{\beta,K}+R_{\gamma,K}). \quad (27)$$

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 comparisons with the functional linear model (FLM) (1) for both dense and sparse designs. Each of 400 simulation runs consisted of a sample of $n = 100$ predictor trajectories X_i, with mean function $\mu_X(s) = s + \sin(s)$, $0 \le s \le 10$, and a covariance function derived from two eigenfunctions, $\phi_1(s) = -5^{-1/2}\cos(\pi s/10)$ and $\phi_2(s) = 5^{-1/2}\sin(\pi s/10)$, $0 \le s \le 10$. The corresponding eigenvalues were chosen as $\lambda_1 = 4$, $\lambda_2 = 1$ and $\lambda_k = 0$, $k \ge 3$, the measurement errors in (11) as $\varepsilon_{ij}{\sim}N(0, 0.5^2)$. To study the effect of Gaussianity, which is of interest especially for the sparse design case, we considered two settings: (i) $\xi_{ik} \sim \mathcal{N}(0, \lambda_k)$ (Gaussian); (ii) ξ_{ik} are generated from the mixture of two normals, $\mathcal{N}\{(\lambda_k/2)^{1/2}, \lambda_k/2)$ with probability $1/2$ and $\mathcal{N}\{-(\lambda_k/2)^{1/2}, \lambda_k/2\}$ with probability $1/2$ (mixture distribution). The number of measurements where each trajectory was sampled was selected from $\{30, \ldots, 34\}$, respectively from $\{5, \ldots, 9\}$ with equal probability for dense respectively sparse cases.

The response variables were generated from the regression model $Y_i = m(\xi_{i1}, \xi_{i2}, \ldots) + \epsilon_i$, where $m(\cdot)$ characterizes functional relationship of the predictor trajectories X_i with the responses Y_i, with errors $\epsilon_i \sim N(0, 0.1)$, so that $\mu_Y = 0$ and $\sigma_Y^2 = 0.1$. We compared the performance of functional quadratic and linear models under two situations: (a) quadratic with

$m(\xi_1, \xi_2, \ldots) = \xi_1 + \xi_2 + \xi_1^2 + \xi_2^2 + \xi_1\xi_2$; (b) linear with $m(\xi_1, \xi_2, \ldots) = 2\xi_1 + \xi_2/2$. The functional quadratic model was implemented as described in Section 3, with number of model components for the predictor processes automatically selected as the smallest number explaining at least 90% of total variation. To evaluate the prediction of new responses for future subjects, we generated 100 new predictor trajectories and responses X_i^* and Y_i^*, respectively, with measurements U_{ij}^* taken at the same time points as U_{ij}. The performance measure tabulated in Table 1 is relative prediction error RPE $= \sum_i (Y_i^* - \hat{Y}_i^*)^2 / Y_i^{*2}$. The results suggest that functional quadratic regression leads to similar prediction errors as functional linear regression when the underlying regression function $m(\cdot)$ is linear, while functional quadratic regression is associated with dramatically improved errors across all scenarios when the underlying function is quadratic.

6. Application to Spectral Data

To demonstrate the usefulness of the proposed methods, we explore the functional relationship between absorbance trajectories (longitudinally measured functional predictor) and the fat content (response) of meat. The data are recorded on a Tecator Infratec Food and Feed Analyzer, a spectrometer that works in the wavelength range 850 - 1050 nm. Each of the $n = 199$ included sampling units consists of the set of measurements made on a specific finely chopped pure meat specimen (Borggaard & Thodberg, 1992). These measurements include the fat content, obtained analytically, and a 100 channel spectrum of absorbances, obtained as log transform of the transmittance measured by the spectrometer, yielding 100 equidistant points at which the spectrum is recorded, giving rise to a dense design. 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 (predictor trajectories) displayed in Figure 1 indicates that the predictor trajectories are smooth. The estimated mean function is in the left panel of Figure 2. Four eigenfunctions are selected for modeling which explain more than 99.8% of the total variation (right panel of Figure 2). The estimated univariate linear function $\beta(t)$ and the bivariate quadratic surface $\gamma(s, t)$ in Figure 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 Figure 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 (20) in dependence on the number of included components in Table 2 demonstrate that for more than three components (as required for reasonably good prediction) the error of functional quadratic model is consistently smaller than that for partial least squares which in turn is smaller than that of functional linear regression.

Acknowledgements

We wish to thank two referees for constructive comments. This research was supported by the National Science Foundation and a NSERC Discovery grant.

Appendix

A1. Estimation Procedures

Using the notation introduced in Section 3, the PACE estimates (best linear estimates) (Yao et al., 2005a) of ξ_k^*, conditional on the observations, are given by

$$\hat{\xi}_k^{*P} = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_k^{*T} \widehat{\Sigma}_{U^*}^{-1} (\boldsymbol{U}^* - \boldsymbol{\mu}_X^*), \qquad k = 1, \ldots, K, \tag{28}$$

where $\boldsymbol{\mu}_X^* = (\mu_X(T_1^*), \ldots, \mu_X(T_m^*))^T$. It follows from results in Müller (2005) that, as designs become dense, $\hat{\xi}_k^{*I}$ (17) and $\hat{\xi}_k^{*P}$ (28) can be considered asymptotically equivalent. Smoothing kernels κ_1, κ_2 are compactly supported smooth densities with zero means and finite variances and are implemented with suitable bandwidth sequences b and h. With $f\{\boldsymbol{\theta}, (s,t), (T_{ij}, T_{il})\} = \theta_0 + \theta_{11}(s - T_{ij}) + \theta_{12}(t - T_{il})$, the one- and two-dimensional smoothers to estimate $C_1(t)$ and $C_2(s,t)$ in (12) are obtained by minimizing

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

$$\sum_{i=1}^{n}\sum_{1\le j\ne l\le N_i}\kappa_2(\frac{T_{ij}-s}{h},\frac{T_{il}-t}{h})\Big[C_i^{(2)}(T_{ij},T_{il})-f\{\boldsymbol{\theta},(s,t),(T_{ij},T_{il})\}\Big]^2, \quad (29)$$

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

$$BIC(K)\propto\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left[-\frac{1}{2\hat\sigma^2}\Big\{U_{ij}-\hat\mu(T_{ij})-\sum_{k=1}^{K}\hat\xi_{ik}\hat\phi_k(T_{ij})\Big\}^2\right]+K\log(\sum_{i=1}^{n}N_i). \quad (30)$$

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 ν_1 and ν_2 be positive integers.

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

Let $b^*=b^*(n)$, $h^*=h^*(n)$, $\tilde h=\tilde h(n)$ denote the bandwidths for estimating $\hat\mu_X$ (26), $\widehat{G}$ (27) and $\hat\sigma$ (2) in Yao et al. (2005a). The Fourier transforms of κ_1 and κ_2 are given by $\kappa_1^F(u)=\int\exp(-iut)\kappa_1(t)dt$ and $\kappa_2^F(u,v)=\int exp(-(iut+ivs)\kappa_2(s,t)ds\,dt$ respectively. The following assumptions are needed for both fixed and random designs,

(A2.1) $\max(b^*,h^*,\tilde h,b,h)\to0$, $\min(nb^{*4},n\tilde h^4,nb^4)\to\infty$, $\max(nb^{*6},n\tilde h^6,nb^6)<\infty$, $\min(nh^{*6},nh^6)\to\infty$, $\max(nh^{*8},nh^8)<\infty$, as $n\to\infty$. Furthermore, assume that $h=O\{\min(b^{1/2},b_1^{1/2},\tilde h^{1/2},h^{*1/2})\}$ for simplicity.

(A2.2) κ_1^F and κ_2^F are absolutely integrable, $\int|\kappa_1^F(u)|du<\infty$, $\int\int|\kappa_2^F(u,v)|dudv<\infty$.

(A2.3) $n^{-1/2}h^{-2}\sum_{k=1}^{\widetilde K}\sum_{\ell=1}^{k}(\pi_k+\pi_\ell)(\lambda_k\lambda_\ell)^{-1}\to0$, as $n\to\infty$, where

$$D_X=\int_{\mathcal{T}^2}\{\widehat{G}(s,t)-G(s,t)\}^2dsdt, \qquad \delta_k=\min_{1\le j\le k}(\lambda_j-\lambda_{j+1}),$$
$$\widetilde K=\inf\{j\ge1:\lambda_j-\lambda_{j+1}\le2D_X\}-1,\ \ \pi_k=1/\lambda_k+1/\delta_k. \quad (31)$$

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 $a_0\le T_{(1)}\le\ldots\le T_{(\widetilde N)}\le b_0$, and let $\Delta=\max\{T_{(m)}-T_{(m-1)}:m=1,\ldots,\widetilde N+1\}$, where $\widetilde N=\sum_{i=1}^nN_i$, $\mathcal{T}=[a_0,b_0]$, $t_{(0)}=$

a_0, and $t_{(N+1)} = b_0$. For the ith subject, suppose that the time points T_{ij} have been ordered non-decreasingly. 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 $\bar{N} = \widetilde{N}/n$. 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 ith 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 following 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) $\Delta = O(\min\{n^{-1/2} b^{*-1}, n^{-1/2} \tilde{h}^{-1}, n^{-1/2} b^{-1}, n^{-1/4} h^{*-1}, n^{-1/4} h^{-1}\})$, $\quad \Delta^* = O(1/\bar{N})$, and $C_1 \bar{N} \le N_{\min} \le N_{\max} \le C_2 \bar{N}$ for some $C_1, C_2 > 0$, $\sup_{t \in \mathcal{T}} E\{U^4(t)\} < \infty$.

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

(A3.3) $\tilde{K}^2 = o(\bar{N}^{1/2})$, $\max_{k \le \widetilde{K}} \|\phi_k'\|_\infty = O(\bar{N}^{1/2})$; $E(\|X'\|_\infty) < \infty$, $E(\|X'^2\|_\infty^2) = o(\bar{N})$.

(A3.4) $\tilde{K}^2 = o(N^{*1/2})$, $\max_{k \le \widetilde{K}} \|\phi_k'\|_\infty = O(N^{*1/2})$, where $\widetilde{K}$ is as in (31).

For sparse designs, denote the marginal and joint densities of T, (T, U) and (T_1, T_2, U_1, U_2) by $g_T(t)$, $g_U(t, u)$, $g_U^*(t_1, t_2, u_1, u_2)$ $g(t, u)$ and $g_2(t_1, t_2, u_1, u_2)$. 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.

(A4.1) (T_{ij}, U_{ij}) and (T_j^*, U_j^*) are distributed as (T, U), T_{ij}, T_j^* are distributed as T ($i = 1, \ldots, n$, $j = 1, \ldots, N_i$). The random variables $N_1, \ldots, N_n$, are distributed as $N^* \sim N$ and are independent of all other random variables. The centered fourth moment is finite, i.e., $E[\{U - \mu_X(T)\}^4] < \infty$.

(A4.2) $g_T(t) > 0$ and $(d^\ell/dt^\ell) g_T(t)$ is continuous on $\mathcal{T}$; $(d^2/dt^2) g_U(t;u)$ is uniformly continuous on $\mathcal{T} \times \Re$; $\{d^2/(dt_1^{\ell_1} dt_2^{\ell_2})\} g_U^*(t_1, t_2, u_1, u_2)$ is uniformly continuous on $\mathcal{T}^2 \times \Re^2$, for $\ell_1 + \ell_2 = 2$, $0 \le \ell_1, \ell_2 \le 2$.

(A4.3) The predictor process $X \in L^2(\mathcal{T})$ is a Gaussian process.

Weak uniform convergence rates for estimated mean $\hat{\mu}_X$ and covariance functions $\widehat{G}$ for sparse designs were derived in Yao et al. (2005a) and analogous arguments apply to $\widehat{C}_1$ and $\widehat{C}_2$ in

(29). Minor modifications of these results yield analogous results for fixed designs. Combining Theorem 1 of Hall & Hosseini-Nasab (2006) and arguments used in the proof of Theorem 2 of Yao et al. (2005a) sets the stage for the proofs of the main results.

LEMMA 1. *Under (A2), (A3.1) and (A3.2) for dense designs or under (A2), (A4.1) and (A4.2) for sparse designs, it holds that*

$$\sup_{t\in\mathcal{T}}|\hat{\mu}_X(t)-\mu_X(t)| = O_p(\frac{1}{n^{1/2}b^*}),\ \sup_{s,t\in\mathcal{T}}|\widehat{G}(s,t)-G(s,t)| = O_p(\frac{1}{n^{1/2}h^{*2}}),$$
$$\sup_{t\in\mathcal{T}}|\widehat{C}_1(t)-C_1(t)| = O_p(\frac{1}{n^{1/2}b}),\ \sup_{s,t\in\mathcal{T}}|\widehat{C}_2(s,t)-C_2(s,t)| = O_p(\frac{1}{n^{1/2}h^2}), \qquad (32)$$

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

$$pr(\sup_{1\le k\le \widetilde{K}}|\hat{\lambda}_k-\lambda_k|\le D_X)=1,\ \sup_{t\in\mathcal{T}}|\hat{\phi}_k(t)-\phi_k(t)| = O_p(\frac{\pi_k}{n^{1/2}h^{*2}}),\quad k=1,\ldots,\widetilde{K}. \qquad (33)$$

LEMMA 2. *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* $K\le\widetilde{K}$ *that*

$$\hat{\beta}_k-\beta_k = O_p(\frac{\pi_k}{n^{1/2}h^2\lambda_k}),\ \hat{\gamma}_{k\ell}-\gamma_{k\ell} = O_p(\frac{\pi_k+\pi_\ell}{n^{1/2}h^2\lambda_k\lambda_\ell}),\quad 1\le\ell<k\le K, \qquad (34)$$
$$\hat{\gamma}_{kk}^D-\gamma_{kk} = O_p(\frac{\pi_k}{n^{1/2}h^2\lambda_k^2}+\frac{1}{\bar{N}^{1/2}}),\ \hat{\gamma}_{kk}^S-\gamma_{kk} = O_p(\frac{\pi_k}{n^{1/2}h^2\lambda_k^2})\quad 1\le k\le K, \qquad (35)$$

where $\hat{\beta}_k$, $\hat{\gamma}_{k\ell}$ *are as in (14) for* $\ell<k$, *and* $\hat{\gamma}_{kk}^D$, $\hat{\gamma}_{kk}^S$ *are as in (15).*

Proof of Lemma 2. It is easy to show the rate for $\hat{\beta}_k$, by observing that $\lambda_k^{-1}<\pi_k$ from (31) 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 1. 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 $(\hat{\lambda}_k\hat{\lambda}_\ell)^{-1}-(\lambda_k\lambda_\ell) = O_p\{n^{-1/2}h^{*-2}\lambda_k^{-1}\lambda_\ell^{-1}(\lambda_k+\lambda_\ell)^{-1}\}$ from Lemma 1 and again $\lambda_k^{-1}<\pi_k$, leading to (34). For (35), the rate of $\hat{\gamma}_{kk}^S$ can be obtained by the same argument used for $\hat{\gamma}_{k\ell}$ with $\ell<k$. To characterize the rate of $\hat{\gamma}_{kk}^D$, it is necessary to quantify $(\hat{\xi}_{ik}^I-\xi_{ik})$, where $\hat{\xi}_{ik}^I$ is as in (17) for the ith subject. Using similar arguments as in the proof of Theorem 1 in Yao & Lee (2006), one finds, with Lemma 1 and (A3.3), that $n^{-1}(\sum_{i=1}^n \hat{\xi}_{ik}^{I4})-E(\xi_k^4) = O_p\{\pi_k/(n^{1/2}h^{*2})+\bar{N}^{-1/2}\}$. Analogous arguments for $\gamma_{k\ell}$, $\ell<k$, complete the proof. □

Proof of Theorem 1. Results (23) and (24) follow immediately from the representations in (10), (14) and Lemma 2. One can easily show (22) by applying the Cauchy-Schwarz inequality to $\sum_{k=K+1}^{\infty}\gamma_{kk}\lambda_k$ and noting that, as $n\to\infty$, $\sum_{k=K+1}^{\infty}\lambda_k^2$ converges even faster to 0 than $\sum_{k=K+1}^{\infty}\lambda_k$. □

Proof of Theorem 2. We first show (26) concerning dense designs. To quantify the prediction error of $(\hat{\xi}_k^{*I}-\xi_k^*)$, where $\hat{\xi}_k^{*I}$ is as in (17), similar arguments as in the proof of Theorem 1 in Yao & Lee (2006) imply that $\hat{\xi}_k^{*I}-\xi_k^*=O_p\{\pi_k/(n^{1/2}h^{*2})+N^{*-1/2}\}$ and analogously $\hat{\xi}_k^{*I}\hat{\xi}_\ell^{*I}-\xi_k^*\xi_\ell^*=O_p\{(\pi_k+\pi_\ell)/(n^{1/2}h^{*2})+N^{*-1/2}\}$. Applying Lemma 2, the discrepancy between $\widehat{Y}^*$ and $Y_K^*=\sum_{k=1}^K\beta_k\xi_k^*+\sum_{1\le\ell\le k\le K}\gamma_{k\ell}\xi_k^*\xi_\ell^*$ is seen to be $O_p\big\{(n^{1/2}h^2)^{-1}\sum_{1\le\ell\le k\le K}(\pi_k+\pi_\ell)(\lambda_k+\lambda_\ell)^{-1}+K^2N^{*-1/2}\big\}$. To find the bound for the remainder term $Q_K=E(Y^*|X^*)-Y_K^*=\sum_{k=K+1}^{\infty}\beta_k\xi_k^*+\sum_{k=K+1}^{\infty}\sum_{\ell=1}^{k}\gamma_{k\ell}\xi_k^*\xi_\ell^*$, it is sufficient to show $E(Q_K^2)=o(R_{\beta,K}^2+R_{\gamma,K}^2)$, where $R_{\beta,K}$, $R_{\gamma,K}$ are as in (21). Since $E(Q_K)=\sum_{k=K+1}^{\infty}\gamma_{kk}\lambda_k$, with the Cauchy-Schwarz inequality and (A1)

$$
\begin{aligned}
E(Q_K^2) &= \mathrm{var}\Big(\sum_{k=K+1}^{\infty}\beta_k\xi_k^*+\sum_{k=K+1}^{\infty}\sum_{\ell=1}^{k}\gamma_{k\ell}\xi_k^*\xi_\ell^*\Big)+\Big(\sum_{k=K+1}^{\infty}\gamma_{kk}\lambda_k\Big)^2\\
&\le \sum_{k=K+1}^{\infty}\beta_k^2\lambda_k+\sum_{k=K+1}^{\infty}\sum_{\ell=1}^{k-1}\gamma_{k\ell}^2\lambda_k\lambda_\ell+\sum_{k=K+1}^{\infty}\gamma_{kk}^2\tau_k+\Big(\sum_{k=K+1}^{\infty}\gamma_{kk}^2\Big)\Big(\sum_{k=K+1}^{\infty}\lambda_k^2\Big), \qquad (36)
\end{aligned}
$$

which implies (26). For the case of sparse designs, $\sup_{t\in\mathcal{T}}|\hat{\lambda}_k\hat{\phi}_k(t)-\lambda_k\phi_k(t)|=O_p\{\pi_k/(n^{1/2}h^{*2})\}$ according to the proof of Theorem 2 in Yao et al. (2005a). For $\tilde{\xi}_k^*=E(\xi_k^*|\boldsymbol{U}^*)$ and $\widetilde{\xi_k^*\xi_\ell^*}=E(\xi_k^*\xi_\ell^*|\boldsymbol{U}^*)$, it is easy to verify that $\hat{\xi}_k^{*P}-\tilde{\xi}_k^*=O_p\{\pi_k/(n^{1/2}h^{*2})\}$ and $\widehat{\xi_k^*\xi_\ell^*}^P-\widetilde{\xi_k^*\xi_\ell^*}=O_p\{(\pi_k+\pi_\ell)/(n^{1/2}h^{*2})\}$. For the truncated quantity $\widetilde{Y}_K^*=\sum_{k=1}^K\beta_k\tilde{\xi}_k^*+\sum_{1\le\ell\le k\le K}\gamma_{k\ell}\widetilde{\xi_k^*\xi_\ell^*}$, we find $\widehat{Y}^*-\widetilde{Y}_K^*=O_p\big\{(n^{1/2}h^2)^{-1}\sum_{1\le\ell\le k\le K}(\pi_k+\pi_\ell)(\lambda_k+\lambda_\ell)^{-1}\big\}$. Considering $\widetilde{Y}^*-\widetilde{Y}_K^*=E\big(\sum_{k=K+1}^{\infty}\beta_k\xi_k^*+\sum_{k=K+1}^{\infty}\sum_{\ell=1}^{k}\gamma_{k\ell}\xi_k^*\xi_\ell^*|\boldsymbol{U}^*\big)$, it suffices to show that $E\{(\widetilde{Y}^*-\widetilde{Y}_K^*)^2\}=o(R_{\beta,K}^2+R_{\gamma,K}^2)$, where the expectation is unconditional (with respect to both $\boldsymbol{U}^*$ and X^*). Noting that $E\{(\widetilde{Y}^*-\widetilde{Y}_K)^2\}\le\mathrm{var}(\widetilde{Y}^*-\widetilde{Y}_K^*)+(\sum_{k=K+1}^{\infty}\gamma_{kk}\lambda_k)^2$, $E\big[\mathrm{var}\big\{E(Y^*-Y_K^*|X^*)|\boldsymbol{U}^*\big\}\big]\ge 0$ and (A3.3), one obtains $\mathrm{var}(\widetilde{Y}^*-\widetilde{Y}_K^*)\le\mathrm{var}\big\{E(Y^*-Y_K|X^*)\big\}$. An upper bound for the right hand side has been obtained in (36), as (A4.3) implies (A1). This completes the proof of (27). □

REFERENCES

BORGGAARD, C. & THODBERG, H. (1992). Optimal minimal neural interpretation of spectra. *Analytical Chemistry* **64**, 545–551.

CAI, T. & HALL, P. (2006). Prediction in functional linear regression. *The Annals of Statistics* **34**, 2159–2179.

CARDOT, H., FERRATY, F., MAS, A. & SARDA, P. (2003). Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics. Theory and Applications* **30**, 241–255.

CARDOT, H. & SARDA, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis* **92**, 24–41.

CASTRO, P. E., LAWTON, W. H. & SYLVESTRE, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28**, 329–337.

FARAWAY, J. J. (1997). Regression analysis for a functional response. *Technometrics* **39**, 254–261.

HALL, P. & HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics* **35**, 70–91.

HALL, P. & HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 109–126.

HE, G., MÜLLER, H. G. & WANG, J. L. (2000). Extending correlation and regression from multivariate to functional data. In *Asymptotics in Statistics and Probability*, M. L. Puri, ed. VSP International Science Publishers, pp. 301–315.

IZEM, R. & KINGSOLVER, J. (2005). Variation in continuous reaction norms: Quantifying directions of biological interest. *American Naturalist* **166**, 277–289.

JAMES, G. M., HASTIE, T. J. & SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.

JONES, M. C. & RICE, J. A. (1992). Displaying the important features of large collections of similar curves. *The American Statistician* **46**, 140–145.

KIRKPATRICK, M. & HECKMAN, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology* **27**, 429–450.

LIU, B. & MÜLLER, H. (2009). Estimating derivatives for samples of sparsely observed functions, with application to on-line auction dynamics. *Journal of the American Statistical Association* .

MELAS, V., PEPELYSHEV, A. & CHENG, R. (2003). Designs for estimating an extremal point of quadratic regression models in a hyperball. *Metrika* **58**, 193–208.

MORRIS, J. S. & CARROLL, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 179–199.

MÜLLER, H.-G. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics. Theory and Applications* **32**, 223–240.

MÜLLER, H.-G. & ZHANG, Y. (2005). Time-varying functional regression for predicting remaining lifetime distributions from longitudinal trajectories. *Biometrics* **61**, 1064–1075.

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

RAMSAY, J. O. & SILVERMAN, B. W. (2002). *Applied functional data analysis*. Springer Series in Statistics. New York: Springer-Verlag.

RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional data analysis*. Springer Series in Statistics. New York: Springer, 2nd ed.

RICE, J. A. (2004). Functional and longitudinal data analysis: Perspectives on smoothing. *Statistica Sinica* , 631–647.

RICE, J. A. & SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **53**, 233–243.

SHEN, Q. & FARAWAY, J. (2004). An F test for linear models with functional responses. *Statistica Sinica* **14**, 1239–1257.

XU, L., JIANG, J.-H., WU, H.-L. & YU, G.-L. S. R.-Q. (2007). Variable-weighted pls. *Chemometrics and Intelligent Laboratory Systems* **85**, 140–143.

YAO, F. & LEE, T. C. M. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society, Series B* **68**, 3–25.

YAO, F., MÜLLER, H.-G. & WANG, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

YAO, F., MÜLLER, H. G. & WANG, J. L. (2005b). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

YAO, F., MÜLLER, H.-G. & WANG, J.-L. (2005c). Functional linear regression analysis for longitudinal data. *The Annals of Statistics* **33**, 2873–2903.

ZHAO, X., MARRON, J. S. & WELLS, M. T. (2004). The functional data analysis view of longitudinal data. *Statistica Sinica* **14**, 789–808.

[*Received June* 2008. *Revised April* 2009]

Table 1. *Monte Carlo estimates of the 25th, 50th and 75th percentiles of relative prediction error, comparing predictions obtained by the Functional Quadratic Model (FQM) and Functional Linear Model (FLM) 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.*

Design	True	Fitted	Gaussian			Mixture		
			25th	50th	75th	25th	50th	75th
Dense	Quadratic	FQM	0.037	0.201	1.403	0.033	0.166	0.950
		FLM	0.204	1.207	30.02	0.162	0.860	17.39
	Linear	FQM	0.047	0.228	1.168	0.031	0.151	0.808
		FLM	0.049	0.269	1.573	0.029	0.162	0.946
Sparse	Quadratic	FQM	0.067	0.358	3.950	0.040	0.207	1.341
		FLM	0.233	1.340	29.16	0.157	0.845	16.01
	Linear	FQM	0.032	0.167	0.934	0.032	0.159	0.840
		FLM	0.038	0.193	1.063	0.037	0.178	0.936

Table 2. *Medians of cross-validated relative prediction errors (PE$\times 10^4$) and of quasi-R^2 (20) ($R_Q^2 \times 100$), for varying numbers of components, comparing functional quadratic model (FQM), functional linear model (FLM) and Partial Least Squares (PLS) for the 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

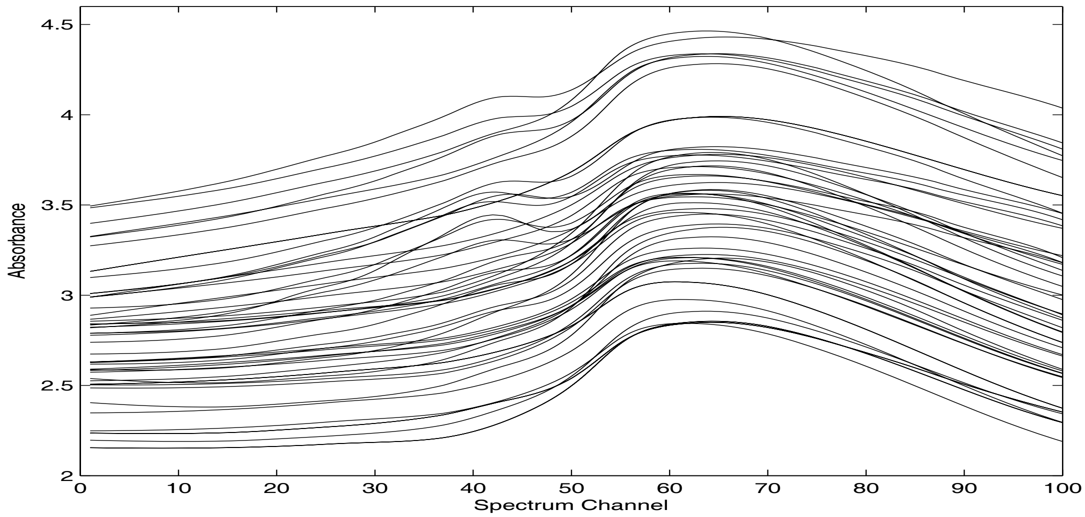

Fig. 1. The functional predictor trajectories, consisting of 100 channel spectra of log-transformed absorbances, for 50 randomly selected meat specimen.

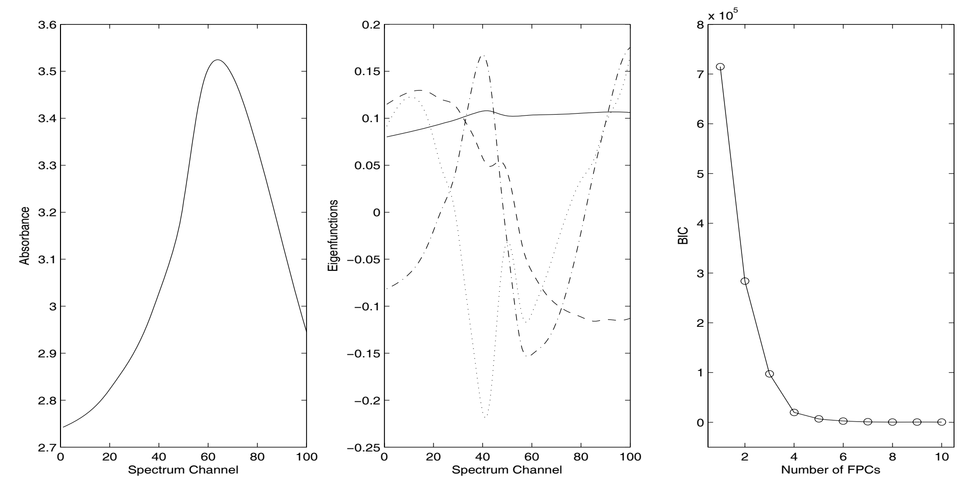

Fig. 2. Smooth estimates of the predictor mean function (left panel) and first (solid), second (dashed), third (dashed-dot) and fourth (dotted) eigenfunctions (right panel), as well as the values of the Bayesian Information Criterion (30), plotted against number of included principal components (right panel).

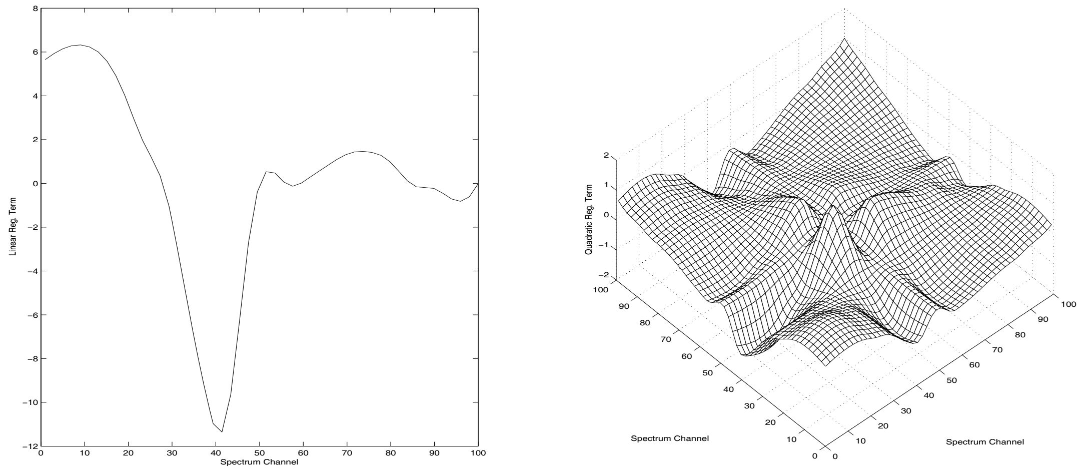

Fig. 3. Estimated parameter functions of functional quadratic regression, with the linear part $\beta(t)$ (left panel) , and the quadratic part $\gamma(s,t)$ (right panel) for spectral data.

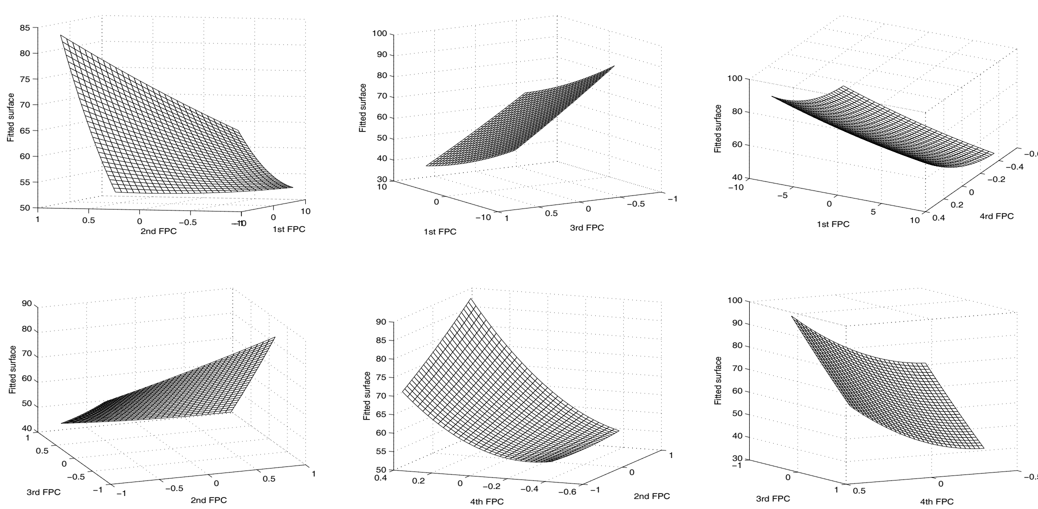

Fig. 4. Fitted quadratic response surface for $\widehat{Y}$ (fat content), obtained by varying two predictor principal components at a time and fixing the other two at 0. Arranged from top left to right and then bottom left to right: $\widehat{Y}$ versus $(\xi_1, \xi_2, 0, 0)$, $(\xi_1, 0, \xi_3, 0)$, $(\xi_1, 0, 0, \xi_4)$, $(0, \xi_2, \xi_3, 0)$, $(0, \xi_2, 0, \xi_4)$ and $(0, 0, \xi_3, \xi_4)$, respectively.