Generalized Functional Linear Models with Semiparametric Single-Index Interactions

Yehua Li,
Department of Statistics, University of Georgia, Athens, GA 30602, yehuali@uga.edu

Naisyin Wang, and
Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107, nwangaa@umich.edu

Raymond J. Carroll[1]
Department of Statistics, Texas A&M University, TAMU 3143, College Station, TX 77843-3143, carroll@stat.tamu.edu

Abstract

We introduce a new class of functional generalized linear models, where the response is a scalar and some of the covariates are functional. We assume that the response depends on multiple covariates, a finite number of latent features in the functional predictor, and interaction between the two. To achieve parsimony, the interaction between the multiple covariates and the functional predictor is modeled semiparametrically with a single-index structure. We propose a two step estimation procedure based on local estimating equations, and investigate two situations: (a) when the basis functions are pre-determined, e.g., Fourier or wavelet basis functions and the functional features of interest are known; and (b) when the basis functions are data driven, such as with functional principal components. Asymptotic properties are developed. Notably, we show that when the functional features are data driven, the parameter estimates have an increased asymptotic variance, due to the estimation error of the basis functions. Our methods are illustrated with a simulation study and applied to an empirical data set, where a previously unknown interaction is detected. Technical proofs of our theoretical results are provided in the online supplemental materials.

Keywords

Functional data analysis; Generalized linear models; Interactions; Latent variables; Local estimating equation; Principal components; Single index models

1 Introduction

In this paper, we are interested in functional data analysis for regression problems with a scalar response and where some of the covariates are functional. The most important model for such problems is the functional generalized linear model, where the response depends on

a linear projection of the functional predictor. Some recent work on this problem includes Ramsay and Silverman (2005), Müller and Stadtmüller (2005), Yao, Müller and Wang (2005), Cardot and Sarda (2005), Cai and Hall (2006), Ferraty and Vieu (2006), Crainiceanu, Staicu and Di (2008) and Di, Crainiceanu, Caffo and Punjabi (2009), among many others.

However, most existing work does not readily accommodate the existence of an interaction between the functional predictor and other covariates. Our paper addresses this issue by introducing a novel class of models that accommodate such an interaction. A special feature of our work, which differs from the existing literature, is that we use a semiparametric single index structure to model the potential interaction between the functional predictor and other covariates. This single index structure not only gives flexibility in describing a complex interaction when there are two or more other covariates but it also allows us to propose a practically feasible two-stage procedure to address estimation and inference issues.

Our new *generalized functional linear models with semiparametric interactions* have the structure that the response depends on latent features of the functional data and their interaction with other possibly multivariate covariates. The "features" in functional data that we are considering are the projections of the functional data onto orthonormal basis functions. These basis functions can be either fixed, e.g., Fourier or wavelet basis functions, or data driven, e.g., principal components. The interaction is modeled semiparametrically, with the regression coefficient function for the functional predictor depending on a single-index function of the multivariate covariates.

After defining the modeling framework in Section 2, we split our study into two tracks. In the first track, in Section 3, we consider the situation that the functional predictor is fully observed and the basis functions are pre-determined, as might occur for Fourier basis functions or some versions of splines. Consequently, the score of the latent features can be evaluated. In this context we propose a backfitting estimation procedure based on local estimating equations. The procedure we propose consists of two stages, which are based on the philosophy of minimum average variance estimation (MAVE, Xia, Tong, Li and Zhu, 2002): in the first stage, we use multivariate kernel weights in the local estimating equations to get consistent initial values; in the second stage, we switch to univariate kernel methods to achieve more efficient estimators. We also derive the asymptotic properties of the methodology.

The second track of the paper, in Section 4, considers the case that the functional features/basis functions are data driven and need to be estimated. At present, the most common dimension reduction device in functional data analysis is functional principal component analysis (FPCA). Some recent work, including Yao, et al. (2005), Hall and Hosseini-Nasab (2006) and Hall, Müller and Wang (2006), have led to a deeper understanding of this method and its properties. We apply the FPCA method based on kernel smoothing as proposed by the authors mentioned above, and then plug the estimated PCA scores into the two stage estimation procedure of Section 3. With the assumption that the number of observations per curve goes to infinity at a sufficient rate, we show that the estimators of the proposed models are still root-n consistent and asymptotically normally distributed. An important finding is that, even when there is sufficient number of observations per curve, the estimation error in FPCA will increase the variance of the final estimator of the functional linear model. This fact is not sufficiently well appreciated in the literature.

We illustrate the numerical performance of our procedure in a simulation study given in Section 5, and by an empirical application to colon carcinogenesis data, where we exhibit a

previously unknown interaction. Final remarks are provided in Section 6, and all proofs are sketched in the Web Appendix, which is available in the online supplemental materials.

2 Model and Data Structure

The data consist of $i = 1, \ldots, n$ independent triples $(Y_i, X_i, \mathbf{Z}_i)$, where Y is the response, X is a longitudinal covariate process and $\mathbf{Z}$ denotes other covariates. Following Carroll, Fan, Gijbels and Wang (1997) and Müller and Stadtmüller (2005), we model the relationship between Y and $\{X(\cdot), \mathbf{Z}\}$ by imposing structures on the conditional mean and variance of Y. The key component of the mean function is a semiparametric functional linear model where the functional coefficient function varies with both t and $\mathbf{Z}$. More precisely, if $E(Y \mid X, \mathbf{Z}) = \mu_Y(X, \mathbf{Z})$, the model is

$$
\begin{aligned}
g\{\mu_Y(X, \mathbf{Z})\} &= \int_{\mathscr{T}} \mathfrak{A}(t, \mathbf{Z}_1) X(t) dt + \beta^{\mathrm{T}} \mathbf{Z}; \\
\mathrm{var}(Y \mid X, \mathbf{Z}) &= \sigma_Y^2 V\{\mu_Y(X, \mathbf{Z})\},
\end{aligned}
\tag{1}
$$

where $g(\cdot)$ and $V(\cdot)$ are known functions, $\mathfrak{A}(\cdot, \mathbf{Z}_1)$ and β are unknown and $\mathscr{T} = [a, b]$ is a fixed interval. We allow the coefficient function $\mathfrak{A}(\cdot, \cdot)$ to depend on a covariate vector $\mathbf{Z}_1$ which is a subset of $\mathbf{Z}$.

Suppose $\psi_1(t), \psi_2(t), \cdots, \psi_p(t)$ are p orthonormal functions on $\mathscr{T}$, and $\xi_j = \int \psi_j(t)\{X(t) - \mu_X(t)\}dt$, for $j = 1, \cdots, p$, where $\mu_X(t) = E\{X(t)\}$. It is commonly assumed that the conditional distribution of Y given $X(\cdot)$ and $\mathbf{Z}$ only depends on $\xi = (\xi_1, \cdots, \xi_p)^{\mathrm{T}}$ and $\mathbf{Z}$. There are two typical structural considerations in functional data analysis, the two tracks of our research. The first track is that of fixed features, where the (ψ_j) are known basis functions, such as Fourier basis functions. Recently Zhou, Huang and Carroll (2008) also showed how to construct an orthonormal basis from B-splines. The second track is of a data-driven nature, e.g., the (ψ_j) are the leading principal components of $X(t)$. The choice of p, as well as the choice of basis functions, is in many cases quite subjective. If we further assume a likelihood function for the data, we can choose p by an AIC criterion, similar to that in Müller and Stadtmüller (2005) and Yao et al. (2005). Because of limited space, such a model selection problem will not be fully explored in this paper.

By assumption, Y depends on $X(\cdot)$ only through the features ξ. Let $\psi(t) = \{\psi_1(t), \cdots, \psi_p(t)\}^{\mathrm{T}}$. To make the model parsimonious without imposing strong parametric structural assumptions, and to allow interactions between X and $\mathbf{Z}_1$, we further assume that

$$
\mathfrak{A}(\cdot, \mathbf{Z}_1) = \psi^{\mathrm{T}}(t)\alpha(\mathbf{Z}_1; \theta), \text{ and that } \alpha(\mathbf{Z}_1; \theta) = \alpha_1 + S(\mathbf{Z}_1; \theta)\alpha_2,
\tag{2}
$$

where $S(\mathbf{Z}_1; \theta)$ is a semiparametric function with a single-index structure described below, and $(\alpha_1; \alpha_2)$ are unknown coefficient vectors. Thus, we assume that the interaction between the longitudinal process $X(t)$ and the multivariate $\mathbf{Z}$ is modeled through a single index function of the multivariate covariate $\mathbf{Z}_1$.

For interpretation of such interaction, we focus on a given $\psi_j(t)$. For this $\psi_j(t)$, the value of ξ_{ij} determines whether the X variable of subject i has elevated values in this direction. The parameters, α_{2j}, θ and $S(\cdot)$, determine how the variable $\mathbf{Z}_i$ interacts with X in this direction through the values of $\alpha_{2j}\xi_{ij}S(\mathbf{Z}_{1i}; \theta)$. When $\mathbf{Z}_1$ is scalar, $S(\mathbf{Z}_1)$ is a nonparametric function. When $\mathbf{Z}_1$ is a $d_1 \times 1$ vector, $S(\mathbf{Z}_1; \theta) = S(\theta^{\mathrm{T}}\mathbf{Z}_1)$, where $\theta = (\theta_1, \cdots, \theta_{d_1})^{\mathrm{T}}$ is a single-index weight vector subject to the usual single-index constraints $\|\theta\| = 1$ and $\theta_1 > 0$. To ensure identifiability, we set the additional constraints that $\|\alpha_2\|^2 = 1$ and $E\{S(\mathbf{Z}_1; \theta)\} = 0$.

Our model is innovative and flexible in the following ways. First, when $S(\mathbf{z}; \boldsymbol{\theta}) \equiv 0$ for all $\mathbf{z}$, model (1) yields the functional generalized linear model of Müller and Stadmüller (2005) as a special case. Second, the nonparametric function $S(\cdot)$ enables us to model nonlinear interactions flexibly while the single-index $\boldsymbol{\theta}^{\mathrm{T}} \mathbf{Z}_1$ allows us to accommodate a multivariate $\mathbf{Z}_1$ without suffering from the curse of dimensionality. Third, the model we propose is not limited to functional data. Given the latent variables $\boldsymbol{\xi}$, the conditional mean of Y can be simplified to

$$g^{-1}\{\mathrm{E}(Y|\boldsymbol{\xi}, \mathbf{Z})\}=\alpha^{\mathrm{T}}(\mathbf{Z}_1;\theta)\boldsymbol{\xi}+\beta^{\mathrm{T}}\mathbf{Z}.$$

(3)

Therefore, our semiparametric interaction model is readily applicable to multivariate data, where $\boldsymbol{\xi}$ is an observed covariate vector.

3 Model Estimation With Fixed Basis Functions

In this section, we consider the case that the basis functions (ψ_j) are known, $X_i(t)$ is fully observed and centered, and consequently $\xi_{ij} = \int X_i(t)\psi_j(t)dt$ is known.

We study this seemingly unrealistic scenario first for two reasons. First, this is a commonly used ideal situation typically considered first in the functional data literature in order to motivate new methodology. A more realistic situation will be studied in our Section 4. Second, as pointed out in the previous section, the semiparametric interaction model we proposed is not limited to functional data. The methods we propose below for this ideal situation are also applicable to the multivariate semiparametric model (3) when $\boldsymbol{\xi}$ is another observed multivariate covariate.

3.1 Local Quasi-likelihood estimation

We estimate $S(\cdot)$ through a local polynomial smoothing approach. Let $\boldsymbol{\Theta}=(\alpha_1^{\mathrm{T}}, \alpha_2^{\mathrm{T}}, \beta^{\mathrm{T}}, \theta^{\mathrm{T}})^{\mathrm{T}}$. Our strategy is to iterate between estimating $\boldsymbol{\Theta}$ while holding $S(\cdot)$ fixed and estimating $S(\cdot)$ via local estimating equations while holding $\boldsymbol{\Theta}$ fixed.

Let $Q(w, y)$ be the quasi-likelihood function satisfying $\partial Q(w, y)/\partial w = (y - w)/V(w)$ (McCullagh and Nelder, 1989). For a given value of $\boldsymbol{\Theta}$, we estimate $S(\upsilon)$ and $S'(\upsilon)$ by the argument $(a_0, a_1)^{\mathrm{T}}$ that minimizes the local quasi-likelihood

$$\sum_{i=1}^{n} Q\left(g^{-1}[\alpha_1^{\mathrm{T}}\boldsymbol{\xi}_i+\{a_0+a_1(\theta^{\mathrm{T}}\mathbf{Z}_{i1} - \upsilon)\}\alpha_2^{\mathrm{T}}\boldsymbol{\xi}_i+\beta^{\mathrm{T}}\mathbf{Z}_i], Y_i\right) w_i(\upsilon),$$

(4)

where $w_i(\upsilon)$ is the local weight for the ith observation.

Model (1) is new, although in various special cases it shares the spirit of previous work on single index models such as Carroll, et al. (1997),Xia et al. (2002) and Xia and Härdle (2006). For computation, we adopt the philosophy of the latter two papers, describing a version of their MAVE method applicable to our problem. Consider $\upsilon = \boldsymbol{\theta}^{\mathrm{T}} \mathbf{Z}_{j1}$ in (4) for each $j \in \{1, \cdots, n\}$ and let a_{0j} and a_{1j} be the estimate of $S(\boldsymbol{\theta}^{\mathrm{T}} \mathbf{Z}_{j1})$ and $S'(\boldsymbol{\theta}^{\mathrm{T}} \mathbf{Z}_{j1})$, respectively.

Summing over $\upsilon = \boldsymbol{\theta}^{\mathrm{T}} \mathbf{Z}_{j1}$ for all j in (4) while holding the values of (a_{0j}, a_{1j}) fixed, and denoting $\mathbf{Z}_{i1} - \mathbf{Z}_{j1}$ by $\mathbf{Z}_{ij,1}$, we can now update the rest of the parameters by minimizing

$$\mathscr{Q}=\sum\nolimits_{j=1}^{n}\sum\nolimits_{i=1}^{n}Q\left[g^{-1}\{\alpha_1^{\mathrm{T}}\boldsymbol{\xi}_i+(a_{0j}+a_{1j}\theta^{\mathrm{T}}\boldsymbol{Z}_{ij,1})\alpha_2^{\mathrm{T}}\boldsymbol{\xi}_i+\beta^{\mathrm{T}}\boldsymbol{Z}_i\},Y_i\right]w_{ij},$$

(5)

with respect to $\boldsymbol{\Theta}$, where w_{ij} are local weights specified in more detail below. As described above, for identifiability purposes we let $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}_2$ satisfy $\|\boldsymbol{\theta}\|=1$ and $\|\boldsymbol{\alpha}_2\|=1$. The constraint $\mathrm{E}\{S(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Z}_1)\}=0$ is replaced by $n^{-1}\sum\nolimits_{j=1}^{n}a_{0j}=0$.

We have now restructured the estimation procedure into one that minimizes the local quasi-likelihood function (5) iteratively with respect to $\{a_{0j},a_{1j},j=1,\cdots,n\}$ and $\boldsymbol{\Theta}$. One remaining task is to specify the local weights w_{ij}. We use two sets of weights. In the initial stage, we let $w_{ij}^{[1]}=H_b(\boldsymbol{Z}_{ij,1})/\sum\nolimits_{\ell=1}^{n}H_b(\boldsymbol{Z}_{\ell j,1})$, where $H(\cdot)$ is a d_1-dimensional kernel function and $H_b(\cdot)=b^{-d_1}H(\cdot/b)$. This will enable us to find a consistent estimator $\tilde{\boldsymbol{\Theta}}$. We then switch to a set of refined weights to gain more efficiency. In the second stage, we carry out the same iteration steps but let $w_{ij}^{[2]}(\check{\theta})=K_h(\check{\theta}^{\mathrm{T}}\boldsymbol{Z}_{ij,1})/\sum\nolimits_{\ell=1}^{n}K_h(\check{\theta}^{\mathrm{T}}\boldsymbol{Z}_{\ell j,1})$, where $K(\cdot)$ is a univariate kernel function, $K_h(\cdot)$ is its scaled version, h is the bandwidth and $\check{\boldsymbol{\theta}}$ is the estimated value of $\boldsymbol{\theta}$ from the previous iteration.

The estimation procedure then is as follows.

Stage 1 (Initial estimator) Iteratively update $S(\cdot)$ and $\boldsymbol{\Theta}$ by minimizing (5) with $w_{ij}=w_{ij}^{[1]}$, until the values converge. At each update of the parameters, we adjust for the constraints, $\|\alpha_2\|=\|\theta\|=1$ and $n^{-1}\sum\nolimits_{i=1}^{n}S(\theta^{\mathrm{T}}\boldsymbol{Z}_{i1})=0$. We denote the converged values as $\tilde{\boldsymbol{\Theta}}$ and $\tilde{S}(\cdot)$. These are consistent estimators which will serve as initial values for Stage 2.

Stage 2 (Refined estimator) To improve efficiency, we replace the initial weights by $w_{ij}=w_{ij}^{[2]}(\tilde{\theta})$, where $\tilde{\boldsymbol{\theta}}$ is the consistent estimator of $\boldsymbol{\theta}$ from Stage 1. The final estimators are denoted as $\hat{\boldsymbol{\Theta}}$ and $\hat{S}(\cdot)$.

In common with the MAVE procedure in Xia and Härdle (2006), the procedure described above does not require a root-n consistent pilot estimator of $\boldsymbol{\Theta}$, which is difficult to obtain in the functional data setting. We will show in Section 3.2 that the initial estimator using the multivariate kernel weights provides a consistent estimator, while the use of the univariate kernel weights with consistently estimated $\boldsymbol{\theta}$ at Stage 2 yields a more efficient asymptotically normally distributed estimator.

In practice, we use a one-step Newton-Raphson update version of the iterative algorithm in Stages 1 and 2, which speeds up the computation considerably. As shown in our proofs for Theorem 1 and 2, the iterations in each stage generate a self-attraction process; that is, asymptotically, the distance between the current and previous estimates of $\boldsymbol{\Theta}$ shrinks to zero along with the iteration. This is one of the standard ways to show convergence properties of an iterative algorithm. It is also used by Xia and Härdle (2006) in a setup simpler than ours. In practice, with n being finite, an algorithm with this asymptotic convergence property could still fail to converge. We have not encountered such numerical difficulties in our study though. We have provided more details about the algorithm in the web supplement.

3.2 Asymptotic Theory

Denote the true parameter as $\boldsymbol{\Theta}_0$, and the true interaction link function as $S_0(\cdot)$. Following the notation in Carroll et al. (1997), let $q_1(x,y)=\{y-g^{-1}(x)\}\rho_1(x)$ and

$q_2(x, y) = \left\{y - g^{-1}(x)\right\} \rho_1'(x) - \rho_2(x)$, where $\rho_\ell(x) = \{dg^{-1}(x)/dx\}^\ell / V\{g^{-1}(x)\}$, $\ell = 1, 2$. As mentioned previously, here we assume the (ξ_{ij}) are known. Estimation of ξ_{ij} and the effect of the estimation error will be discussed in Section 4. We make the following assumptions.

Assumption (C1)

(C1.1) The marginal density of Z_1 is positive and uniformly continuous, with a compact support $\mathscr{D} \in \mathbb{R}^{d_1}$. Let $\mathscr{U} = \theta_0^T Z_1$ be the true single-index, with a density $f_{\mathscr{U}}(u)$. Assume $f_{\mathscr{U}}$ is twice continuously differentiable.

(C1.2) $g^{(3)}(\cdot)$ and $V^{(2)}(\cdot)$ exist and are continuous.

(C1.3) $S_0(\cdot)$ is twice continuously differentiable.

(C1.4) For all x, $\rho_2(x) > 0$ and $d\rho_2(x)/dx$ is bounded.

Assumption (C2)

(C2.1) $H(\cdot)$ is a symmetric d_1 dimensional probability density function, $\int H(\mathbf{x}) \mathbf{x} \mathbf{x}^T d\mathbf{x} = I$. $b \to 0$, $\log(n)/(nb^{d_1+2}) \to 0$.

(C2.2) $V_1(\mathbf{z}; \Theta, S)$, $V_2(\mathbf{z}; \Theta, S)$ and $V_3(\mathbf{z}; \Theta, S)$ defined in Web Appendix B.1 are twice continuously differentiable in $\mathbf{z}$, and Lipschitz continuous in Θ.

Assumption (C3)

(C3.1) $K(\cdot)$ is a symmetric univariate probability density function, with support on $[-1, 1]$, $\sigma_K^2 = \int u^2 K(u) du$, $\kappa_2 = \int K^2(u) du$. $h \sim n^{-\lambda}$, $1/6 < \lambda < 1/4$.

(C3.2) $\mathscr{V}_1(\mathbf{z}; \theta)$, $\mathscr{V}_2(\mathbf{z}; \theta)$ and $\mathscr{V}_3(\mathbf{z}; \theta)$ in Web Appendix B.2 are twice continuously differentiable in $\mathbf{z}$, and Lipschitz continuous in θ.

Here are the two main results in this section.

<u>THEOREM 1</u> *(Consistency of the initial estimator) Under conditions (C1)–(C3)*, $\|\tilde{\Theta} - \Theta_0\| \to 0$ *with probability 1.*

<u>THEOREM 2</u> *(Asymptotic normality of the refined estimator) Under conditions (C1)–(C3)*,

$$\sqrt{n}(\widehat{\Theta} - \Theta_0) \to \text{Normal}(\mathbf{0}, \mathscr{A}^- \mathscr{B} \mathscr{A}^-),$$
$$\sqrt{nh}\{\widehat{S}(u) - S_0(u) - h^2 S_0^{(2)}(u) \sigma_k^2/2\} \to Normal\{0, \sigma_s^2(u)\},$$

where $\mathscr{A}$ and $\mathscr{B}$ are defined in the Appendix,

$\sigma_s^2(u) = \kappa_2 E\{(\alpha_{2,0}^T \xi_i) \epsilon_i^2 | \theta_0^T Z_{i1} = u\} / \{\sigma_v^4(u) f_{\mathscr{U}}(u)\}$, $\sigma_v^2(u) = E\{\rho_2(\eta_i)(\alpha_{2,0}^T \xi_i)^2 | \theta_0^T Z_{i1} = u\}$.

Remark: The algorithm we use is similar to the MAVE method (Xia et al. 2002), which is a relatively new development in the literature on backfitting methods. Unlike many other back-fitting algorithms, the method we are using does not require any artificial undersmoothing for the nonparametric function $S(\cdot)$, see our condition (C3.1) on the rate of the bandwidth. The method also differs from the algorithm in Carroll et al. (1997) in that we do not need a root-n consistent pilot estimator for the parametric components using approaches such as Sliced Inverse Regression (Li, 1991). Instead, the initial estimator in Stage 1 of our procedure will provide a consistent pilot estimator. The refined approach in Stage 2 further results in an efficiency gain which is mainly contributed by a faster convergence rate in the nonparametric estimation at this stage. An efficiency gain due to the

equivalent reason by switching to refined kernel weights in Stage 2 was documented by Xia and Härdle (2006) for partially linear single-index models; see also Section 5.1 where we illustrate this numerically.

4 Estimation With Principal Components

4.1 Background

The most popular dimension reduction method in functional data analysis is principal component analysis, leading to estimated basis functions. Here we focus on the common scenario in which $\psi_j(t)$ denotes the eigenfunction in functional principal component analysis. For the longitudinal process $\{X(t), t \in \mathcal{T}\}$ with mean function $\mu_X(t)$, the covariance

function is $R(s, t) = \text{cov}\{X(s), X(t)\}$ with eigen-decomposition $R(s, t) = \sum_{j=1}^{\infty} \omega_j \psi_j(s) \psi_j(t)$, where the ω_j are the non-negative eigenvalues of $R(\cdot, \cdot)$, which, without loss of generality, satisfy $\omega_1 > \omega_2 > \cdots > 0$, and the ψ_j's are the corresponding eigenfunctions. The Karhunen-Loève expansion of $X(t)$ is

$$X(t) - \mu_X(t) = \sum_{j=1}^{\infty} \xi_j \psi_j(t),$$

(6)

where $\xi_j = \int \psi_j(t)\{X(t) - \mu_X(t)\}dt$ has mean zero, with $\text{cov}(\xi_j, \xi_k) = I(j = k)\omega_j$.

Under this framework, our model (1) means that Y is dependent on the leading p principal components in $X(\cdot)$. The reason that we assume that only a finite number of PC's are relevant is that, in functional data that is commonly seen in biological applications, estimation of high order PC's is highly unstable and difficult to interpret, see the comments in Rice and Silverman (1991) and Hall, et al. (2006). Yao et al. (2005) proposed to use an AIC criterion to choose p. Even though, to the best of our knowledge, there is no theoretical support behind the use of AIC in the existing literature, we found it a very sensible criterion in our numerical investigations which we have reported in Section 5. We made a slight modification of their method in terms of counting the number of parameters, and at least in our simulations were able to select the correct number of PC in every case. More work on this general topic clearly needs to be done.

It often happens that the covariate process we observe contains additional random errors and instead we observe

$$W_{ij} = X_i(t_{ij}) + U_{ij}, \ j = 1, \cdots, m_i,$$

(7)

where U_{ij} are independent zero-mean errors, with $\text{var}\{U_i(t)\} = \sigma_u^2$, and the U_{ij} are also independent of $X_i(\cdot)$ and Z_i.

4.2 Principal Components Analysis via Kernel Smoothing

Many functions and parameters in the expressions given above can be estimated from the data. Thus, $\mu_X(\cdot)$ and $R(\cdot, \cdot)$ can be estimated by local polynomial regression, and then $\psi_k(\cdot)$, ω_k and σ_u^2 can be estimated using the functional principal component analysis method proposed in Yao, et al. (2005) and Hall, et al. (2006). We now briefly describe the method. We first estimate $\mu_X(\cdot)$ by a local linear regression, $\hat{\mu}_X(t) = \hat{a}_0$ where

$$(\widehat{a}_0, \widehat{a}_1) = \mathrm{argmin}_{a0,a1} \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{m_i} \{W_{ij} - a_0 - a_1(t_{ij} - t)\}^2 K\{t_{ij} - t)/h_\mu\},$$

$K(\cdot)$ is a kernel function and h_μ is the bandwidth for estimating μ_X. Letting $\varpi_{XX}(s, t) = \mathrm{E}\{X(s)X(t)\}$, define $\hat{\varpi}_{XX}(s, t) = \hat{a}_0$, where $(\hat{a}_0, \hat{a}_1, \hat{a}_2)$ minimizes

$$\sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{m_i} \sum\nolimits_{k \neq j} \{W_{ij}W_{ik} - a_0 - a_1(t_{ij} - s) - a_2(t_{ik} - t)\}^2 K\left(\frac{t_{ij} - s}{h_\varpi}\right) K\left(\frac{t_{ik} - t}{h_\varpi}\right).$$

Then $\hat{R}(s, t) = \hat{\varpi}_{XX}(s, t) - \hat{\mu}_X(s)\hat{\mu}_X(t)$. The (ω_k) and $\{\psi_k(\cdot)\}$ can be estimated from an eigenvalue decomposition of $\hat{R}(\cdot, \cdot)$ by discretization of the smoothed covariance function, see Rice and Silverman (1991) and Capra and Müller (1997). By realizing that $\sigma_w^2(t) = \mathrm{var}\{W(t)\} = R(t, t) + \sigma_u^2$, we let $\widehat{\sigma}_w^2(t) = \widehat{a}_0 - \widehat{\mu}_x^2(t)$, where $(\hat{a}_0, \hat{a}_1)$ minimizes

$$\sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{m_i} \{W_{ij}^2 - a_0 - a_1(t_{ij} - t)\}^2 K\{t_{ij} - t)/h_\sigma\},$$

and σ_u^2 is estimated by $\widehat{\sigma}_u^2 = (b - a)^{-1} \int_a^b \{\widehat{\sigma}_w^2(t) - \widehat{R}(t, t)\} dt$.

There are two ways to predict the principal component scores, ξ_{ik}. The first is the PACE method proposed by Yao, et al. (2005). By assuming the covariate process $X_i(t)$ and the measurement errors U_{ij} are jointly Gaussian, one can show that the conditional expectation of ξ_{ij} given $W_i = (W_{i1}, \cdots, W_{i,m_i})^\mathrm{T}$ is $\tilde{\xi}_{ik} = \omega_k \psi_{ik} \sum_{W_i}^{-1} (W_i - \mu_{Xi})$, where $\psi_{ik} = \{\psi_k(t_{i1}), \cdots, \psi_k(t_{i,m_i})\}^\mathrm{T}$, $\mu_{Xi} = \{\mu_X(t_{i1}), \cdots, \mu_X(t_{i,m_i})\}^\mathrm{T}$, and $\sum_{W_i} = \mathrm{cov}(W_i) = \{R(t_{ij}, t_{ij'}) + \sigma_u^2 I(t_{ij} = t_{ij'})\}_{j,j'=1}^{m_i}$. The PACE predictors of the principal component scores are calculated by plugging the kernel estimates into the conditional expectation expression,

$$\widehat{\xi}_{ik} = \widehat{\omega}_k \widehat{\psi}_{ik} \widehat{\Sigma}_{W_i}^{-1} (W_i - \widehat{\mu}_{Xi}). \tag{8}$$

An alternative method is by numerical integration. Motivated by the definition that $\xi_{ik} = \int \{X_i(t) - \mu_X(t)\}\psi_k(t)dt$, we define

$$\widehat{\xi}_{ik} = \sum_{j=2}^{m_i} \{W_{ij} - \widehat{\mu}_x(t_{ij})\}\widehat{\psi}_k(t_{ij})(t_{ij} - t_{i,j-1}). \tag{9}$$

In our numerical experience, when we have densely sampled functional data, i.e. m_i is large and the t_{ij}'s are evenly spread in the interval $[a, b]$, the two predictors given in (8) and (9) have comparable performance. The outcomes of the numerical integral predictor, which is easier to use in our theoretical justifications, are reported in Section 5.

4.3 Asymptotic Results for Plugging-in Estimated PC

In this section, we study the asymptotic properties of the estimators proposed in Section 3 when the principal component scores are replaced by their estimates. Without loss of generality, we assume that $\mu_X(t) = 0$, and that there are the same number of discrete

observations for each subject, i.e. $m_i = m$ for all i. Throughout, we assume that $m \to \infty$. We will first focus on the properties of the initial estimators that are obtained at the end of Stage 1. We then provide conditions and the corresponding asymptotic properties of the refined estimators.

4.3.1 Properties of the initial estimators

Assumption (C4)

(**C4.1**)(Hall, et al., 2006) The process $X(\cdot)$ is independent of the measurement error U in
  (7), $E(U) = 0$ and $var(U) = \sigma_u^2$. For some $c > 0$,

  $$\max_{j=0,1,2} E\left\{ \sup_{t \in \mathcal{T}} |X^{(j)}(t)|^c \right\} + E(|U|^c) < \infty.$$

(**C4.2**)We have $\omega_1 > \omega_2 > \cdots > \omega_p > \omega_{p+1} \geq 0$ and ϕ_k is twice continuously differentiable
  on $\mathcal{T}$, for $k = 1, \cdots, p$.

(**C4.3**)The t_{ij} are equally spaced in $\mathcal{T}$, i.e. $t_{ij} = a + (j/m) \times (b - a)$, for $j = 1, \cdots, m$. $m \to$
  ∞.

(**C4.4**)$h_{\varpi} \sim n^{-\lambda_{\varpi}}$, $1/4 < \lambda_{\varpi} < 1/3$.

In Assumption (C4.3) we assume the (t_{ij}) are fixed and equally spaced, thus allowing convenient mathematical derivations. The conclusion of Lemma 1 holds for random designs. The proofs of the following lemma and theorem are provided in the Web Appendix.

<u>LEMMA</u> **1** *Under Assumption (C4), $\|\widehat{\psi}_k - \psi_k\| = O_p\{h_{\varpi}^2 + (nh_{\varpi})^{-1/2}\}$, for $k = 1, \cdots, p$. Let $\hat{\xi}_{ik}$ be the estimated principal component score defined in (9). Then $\hat{\xi}_{ik} - \xi_{ik} = O_p(m^{-1/2} + n^{-1/3})$ for $i = 1, 2, \cdots, n$, and $k = 1, \cdots, p$.*

<u>THEOREM</u> **3** *Suppose $\tilde{\Theta}$ is the initial estimator of Θ defined in Section 3, using the estimated principal component scores. Assume that the bandwidth b for the multivariate kernel satisfies $mb \to \infty$. Under conditions (C1), (C2) and (C4), $\|\tilde{\Theta} - \Theta_0\| \to 0$ with probability 1.*

4.3.2 Properties of the Refined Estimators—In this subsection we will discuss the scenario under which we can pretend that we know the entire trajectory of $X_i(\cdot)$. More precisely, we give conditions that ensure that the smoothing errors are asymptotically negligible. Consequently, we can smooth each curve and effectively eliminate effects from the measurement error in (7). Such a "smooth first, then perform estimation" procedure was considered and justified by Zhang and Chen (2007) in the functional linear model setting.

Let $\tilde{X}_i(t)$ be the estimated trajectory of $X_i(t)$, i.e., the outcome of applying local linear smoothing to W_i. Let h_X be the bandwidth for smoothing each curve. Without loss of generality, assume the $X_i(t)$ have been centered, i.e., the mean curve was subtracted out from each estimated trajectory. One can estimate the covariance function by

$\tilde{R}(s, t) = n^{-1} \sum_{i=1}^{n} \tilde{X}_i(s)\tilde{X}_i(t)$. One can also estimate the eigenfunctions by an eigen-

decomposition of $\tilde{R}$, and the principal component scores by $\widehat{\tilde{\xi}}_{ik} = \int_a^b \tilde{X}_i(t)\widehat{\psi}_k(t)dt$. Zhang and Chen (2007) showed in their Theorem 4 that when sampling points are sufficiently dense on each curve, the principal component analysis method given above is asymptotically equivalent to knowing the entire trajectory of each X_i.

Assumption (C5)

(**C5.1**)$\mathcal{V}(t_1, t_2, t_3, t_4) = E\{X(t_1)X(t_2)X(t_3)X(t_4)\} - R(t_1, t_2)R(t_3, t_4)$ exists for all $t_j \in \mathcal{T}$.

(C5.2) We have $m \geq Cn^{\kappa}$, with $\kappa > 5/4$.

(C5.3) The t_{ij} are independent identically distributed random design points with density function $f(\cdot)$, where f is bounded away from 0 on $\mathscr{T}$ and is continuously differentiable.

(C5.4) $h_X = O(n^{-\kappa/5})$.

<u>LEMMA</u> **2** *Let $\mathbb{Z}(s, t)$ be a zero-mean Gaussian random field on $\{(s, t); s, t \in \mathscr{T}\}$, with covariance function $\mathscr{G}$ defined in Assumption (C5.1). Under assumptions (C4.1)–(C4.2) and (C5), $\|\tilde{X}_i - X_i\| = o_p(n^{-1/2})$, $\sqrt{n}(\tilde{R} - R) \to \mathbb{Z}$, and*

$$\widehat{\psi}_j(t) - \psi_j(t) = n^{-1/2} \sum_{k \neq j} (\omega_j - \omega_k)^{-1} \psi_k(t) \int \mathbb{Z} \psi_j \psi_k + o_p(n^{-1/2}),$$
$$\widehat{\xi}_{ij} - \xi_{ij} = n^{-1/2} \sum_{k \neq j} \xi_{ik} (\omega_j - \omega_k)^{-1} \int \mathbb{Z} \psi_j \psi_k + o_p(n^{-1/2}).$$

The limit distribution of $\tilde{R}$ in Lemma 2 was given in Theorem 4 in Zhang and Chen (2007), whereas the last two equations are direct results of (2.8) of Hall and Hosseini-Nasab (2006).

By plugging in the estimated principal component scores, we have the following asymptotic results for the refined estimator defined in Section 3.

<u>THEOREM</u> **4** *Let $\widehat{\Theta}$ be the refined estimator with plugged in estimated PC scores. Under conditions (C1), (C3), (C4.1)–(C4.2) and (C5),*

$$\sqrt{n}(\widehat{\Theta} - \Theta_0) \to Normal\{\mathbf{0}, \mathscr{A}^-(\mathscr{B} + \mathscr{B}_1)A^-\},$$

where $\mathscr{A}$, $\mathscr{B}$ and $\mathscr{B}_1$ are defined in the Appendix. *In addition, $\hat{S}(\cdot)$ has the same asymptotic distribution as that in Theorem 2.*

By comparing the asymptotic distributions in Theorem 2 and Theorem 4, we can clearly see the additional variation, $\mathscr{B}_1$, which is the consequence of the estimation error in the estimated principal components.

Remarks:

1. Taking the numerical integration method in (9) as an example, there are two sources of error in estimating FPCA scores: the first source is the prediction error for the PC scores given the true eigenfunctions; the second type of error is caused by plugging in the estimated eigenfunctions. Both sources of error tend to be ignored in the functional data literature.

2. One important contribution of our work is that we bring attention to the second source of error mentioned above. Even if we have all the information on each curve, the eigenfunction can only be estimated in a root-n rate (Hall and Hosseini-Nasab, 2006). This estimation error may seem to be small, but it affects PC score estimation in all subjects, and as illustrated in our theory this error will also affect the variance of the final estimator in regression.

3. Plugging in the predicted score is an idea similar to regression calibration (Carroll et al., 2006). With condition (C5.2), we basically assumed that the first type of error (prediction error for the PC scores given the true eigenfunctions) is negligible, which is a typical assumption in the functional data literature, see Müller and Stadtmüller (2005). This allows us to focus all attention on the second source of

errors. Our model considers complex functional interactions and by its nature would need moderate to large sample sizes in order to reach sensible outcomes and reliable findings. When n and/or m are small, whether one could or should consider such a complex model is an interesting yet hard problem, which calls for future research.

4. When condition (C5.2) is violated, the first type of error described above will prevail. Since these prediction errors are independent among subjects, they are similar to the classical error in the measurement error literature and tend to cause bias in the final estimator. Therefore, whether m is sufficiently large can be empirically examined by checking the bias of the final estimator in a simulation study.

5 Numerical Studies

5.1 Logistic Regression Simulation Study

We performed a simulation study to illustrate the numerical performance of the proposed method. Our simulation setting resembles the empirical data described in Section 5.2. We generated the longitudinal process as a Gaussian process with mean function $\mu_X(t) = (t - 0.6)^2 - 0.1$ for $t \in [0, 1]$. The covariance function of the process had 2 principal components, $\psi_1(t) = 1$, $\psi_2(t) = \sqrt{2} \sin(2\pi t)$, and the eigenvalues were $\omega_1 = 1$ and $\omega_2 = 0.6$. There were $m = 30$ discrete observations on each curve, with random observation time points being uniformly distributed on the interval $[0, 1]$. The measurements are contaminated with zero-mean Gaussian error with variance $\sigma_\varepsilon^2 = 0.1$. We generated a binary response Y from the model $\mathrm{pr}(Y = 1 | \boldsymbol{Z}, \boldsymbol{\xi}) = H\{\alpha_1^{\mathrm{T}} \boldsymbol{\xi} + S(\theta^{\mathrm{T}} \boldsymbol{Z}_1) \alpha_2^{\mathrm{T}} \boldsymbol{\xi} + \beta^{\mathrm{T}} \boldsymbol{Z}\}$, where $H(\eta) = 1/\{1 + \exp(\eta)\}$ is the logistic distribution function, $\boldsymbol{Z} = (1, \boldsymbol{Z}_1^{\mathrm{T}}, Z_2)^{\mathrm{T}}$, $\boldsymbol{Z}_1$ is a 2-dimensional random vector with a uniform distribution on $[0, 1]^2$, and Z_2 is a binary variable with $\mathrm{pr}(Z_2 = 1) = 0.5$. We let $\alpha_1 = (0.8, -1)^{\mathrm{T}}$, $\alpha_2 = (1/\sqrt{2}, 1/\sqrt{2})^{\mathrm{T}}$, $\theta = (1/\sqrt{2}, 1/\sqrt{2})^{\mathrm{T}}$, and $\boldsymbol{\beta} = (-1, 2, -2, 2)$. We let $S(\cdot)$ be a sine bump function similar to that used in Carroll et al. (1997), $S(t) = 2\sin\{(t - c_1)/(c_2 - c_1)\}$, where $c_1 = 1/\sqrt{2} - 1.645/\sqrt{12}$ and $c_2 = 1/\sqrt{2} + 1.645/\sqrt{12}$. We then standardized according to the constraints, which resulted in the true $\boldsymbol{\alpha}_1$ being

$$\alpha_1^* = \alpha_1 + \mathrm{E}\{S(\theta^{\mathrm{T}} \boldsymbol{Z}_1)\}\alpha_2.$$

We set $n = 500$ and repeated the simulation 200 times. For each simulated data set we performed FPCA and fit the model using the algorithm described in Section 3, where ξ_{ik} was predicted by the direct numerical integral predictor. We obtained almost identical outcomes using the PACE predictor (not reported). For the iteration procedure, we declared convergence if $\max_k |\tilde{\boldsymbol{\Theta}}_{k,\mathrm{curr}} - \tilde{\boldsymbol{\Theta}}_{k,\mathrm{prev}}|$ is less than a pre-determined value, 10^{-6}, where $\tilde{\boldsymbol{\Theta}}_{k,\mathrm{prev}}$ and $\tilde{\boldsymbol{\Theta}}_{k,\mathrm{curr}}$ denote the previous and current value of the k-th component of $\tilde{\boldsymbol{\Theta}}$. The number of principle component, p, was determined using AIC with the total number of parameters in the penalty equal to the number of PC scores.

The Monte Carlo (MC) means, standard deviations (SD) and biases of the refined estimator are presented in Table 1. All estimates behaved reasonably well. The comparisons of the MC biases and the MC SD's indicate that the proposed estimator has a relatively smaller bias than the standard error. This seems to confirm the theoretical results that they are asymptotically unbiased. For all simulated data sets, the AIC criterion selected $p = 2$ which is the correct number of components. To confirm the efficiency gain of our refined estimator $\hat{\boldsymbol{\Theta}}$ over the initial estimator $\tilde{\boldsymbol{\Theta}}$, we made a comparison of the mean squared error (MSE) of the two estimators. We found that $\mathrm{MSE}(\tilde{\boldsymbol{\Theta}})/\mathrm{MSE}(\hat{\boldsymbol{\Theta}}) = 1.34$. In other words, we have 34% efficiency gain by switching to refined kernel weights.

We compare the proposed approach to two other alternatives. First, we fit a misspecified reduced model without interactions, where $P(Y=1|Z, \xi)=H(\alpha_1^T \xi + \beta^T Z)$. The results are reported in Table 1. An immediate observation is elevated biases in the estimates of α_1 and β due to the misspecification of the model by not considering the interaction.

To assess the variation-inflation effect due to the PC scores being predicted, we also fit the model using the true principal component scores and compare the variances of the final estimators with those in Table 1. In Table 2, we present the relative variance inflation factor for using the true PC scores, which is defined as the ratio of the variance for each parameter when estimating the FPCA scores to the variance when the true FPCA scores are known. One clearly sees that the relative variance inflation factors are uniformly greater than one, with the highest inflation occurring for the coefficients of the second principal component, α_{12} and α_{22}. This phenomena can be explained by the fact that the second PC is much harder to estimate than the first.

5.2 Colon Carcinogenesis Example

5.2.1 Background—The colon carcinogenesis data that we are using are similar to those analyzed in Morris et al. (2001, 2003), Li et al. (2007) and Baladandayuthapani et al. (2008). The biomarker of interest in this experiment is p27, which is a protein that inhibits the cell cycle. We used 12 rats randomly assigned to 2 diet groups (corn oil diet or fish oil diet) and 2 treatment groups (with/without butyrate supplement). Each rat was injected with a carcinogen, and then sacrificed 24 hours after the injection.

Beneath the colon tissue, there are pore structures called 'colonic crypts'. A crypt typically contains 25 to 30 cells, lined up from the bottom to the top. The stem cells are at the bottom of the crypt, where daughter cells are generated. These daughter cells move towards the top as they mature. We sampled about 20 crypts from each of the 12 rats. The p27 expression level was measured for each cell within the sampled crypts. As previously noted in the literature (Morris et al. 2001, 2003), the p27 level measurements in the logarithm scale are natural functional data, since they are indexed by the relative cell location within the crypt. We have $m = 25$–30 observations (cells) on each function. In the literature, it has been noted that there is spatial correlation among the crypts within the same rat (Li et al., 2007, Baladandayuthapani et al., 2008). In this experiment, we sampled crypts sufficiently far apart so that the spatial correlations are negligible, and thus we can assume that the crypts are independent.

In this example, the response Y is the fraction of cells undergoing apoptosis (programmed cell death) in each crypt, to which we applied the usual arcsine-square root transformation. The functional data $X(t)$ is the p27 level of a cell at relative location t. Write $s_0 = 1$, $s_1 =$ indicator of a fish oil diet, $s_2 =$ indicator of butyrate supplementation, and $s_3 =$ mean proliferation level. Then $Z_1=(s_1, s_2, s_3)^T$, $Z=(s_0, Z_1^T)^T$ and $Z^T\beta = \beta_0 + s_1\beta_{\text{fish}} + s_2\beta_{\text{buty}} + s_3\beta_{\text{prolif}}$. Similarly, $S(Z_1, \theta) = S(s_1\theta_{\text{fish}} + s_2\theta_{\text{buty}} + s_3\theta_{\text{prolif}})$.

5.2.2 Initial Model Fits—We first performed FPCA on the p27 data. In Figure 1, we show $\hat{\mu}_X(t)$, $\hat{R}(s, t)$ and the first 3 estimated PCs. The first 3 eigenvalues are 0.871, 0.019 and 0.005 respectively. The estimated variance for the measurement error is $\widehat{\sigma}_\varepsilon^2=0.103$. We again implemented the AIC criterion to choose the number of significant principal components in the p27 data, and $p = 2$ was picked. We thus only include the first 2 principal components of the p27 data in the regression model, so that $\alpha_1 = (\alpha_{11}, \alpha_{12})^T$ and $\alpha_2=(\alpha_{21}, \alpha_{22}^T)$. The estimated parameters are presented in Table 3, and the estimated $S(\cdot)$ is shown in Figure 2.

We estimated the standard error of our parameter estimates by a bootstrap procedure. Of course, after resampling the individual curves (crypts) we reran the functional principal component analysis in the bootstrap sample, so that the bootstrap standard error estimator automatically takes into account the estimation error in FPCA. We then refit the model to the new sample. The procedure was repeated 1, 000 times, and the standard deviations of the estimators in the bootstrap samples were used as our standard error estimators. We then tested each parameter in turn to see whether the true parameter is equal to 0.00. The estimated standard errors and the p-values are reported in Table 3.

To summarize, the parameters, β_{fish}, α_{22}, θ_{buty} and θ_{prolif}, have corresponding p-values less than 0.01. For the parametric main effect, the positive significant β_{fish} implies that the fish oil diet enhances the apoptotic index. The positive $\hat{\beta}_{\text{buty}}$ and the negative $\hat{\beta}_{\text{prolif}}$ suggest that the butyrate supplement and the increasing proliferation level are positively and negatively associated with the apoptotic index, respectively, even though neither is significant at the main effect level. The statistically significant results for α_{22}, θ_{buty} and θ_{prolif} imply that the second principal component of the p27 process is interacting with butyrate supplementation and proliferation.

5.2.3 Testing for Interactions

In addition to testing if each individual parameter is equal to 0.00, it is important to have an overall test of whether interaction really exists in this data set. In our model, this is equivalent to testing the null hypothesis $H_0 : S(\cdot) \equiv 0$. We test this null hypothesis using the following parametric bootstrap procedure:

Step 1. Fit the model under the null hypothesis given as $\text{E}(Y_i | \boldsymbol{\xi}_i, \boldsymbol{Z}_i) = \boldsymbol{\xi}_i^{\text{T}} \alpha_1 + \boldsymbol{Z}_i^{\text{T}} \beta$. Denote the estimators under the reduced model as $\hat{\boldsymbol{\alpha}}_{1,\text{red}}$ and $\hat{\boldsymbol{\beta}}_{\text{red}}$, and call the estimated conditional variance of Y under the reduced model $\widehat{\sigma}^2_{Y,\text{red}}$. We assume the data are Gaussian, and calculate the log likelihood ratio of the full model with interaction to the reduced model above.

Step 2. Use the estimated reduced model as the true one to generate bootstrap samples. We resample with replacement n crypts from the original collection. For the ith resampled crypt, we retain the original covariate vector $\boldsymbol{Z}_i$, use the estimated principal component score and estimated eigenfunction to regenerate the p27 observations

$$W_{ij}^* = \widehat{\mu}_x(t_{ij}) + \sum_{k=1}^{K} \widehat{\xi}_{ik} \widehat{\psi}_k(t_{ij}) + U_{ij}^*, \text{ where } U_{ij}^* = \text{Normal}(0, \widehat{\sigma}^2_u), \widehat{\xi}_{ik}, \widehat{\psi}_k(\cdot) \text{ and } \widehat{\sigma}^2_u \text{ are estimates}$$

from the original data. We choose $K = 2$ to be consistent with our data analysis. Then Y_i^* is generated by $Y_i^* \sim \text{Normal}(\widehat{\boldsymbol{\xi}}_i^{\text{T}} \widehat{\alpha}_{1,\text{red}} + \boldsymbol{Z}_i^{\text{T}} \widehat{\beta}_{\text{red}}, \widehat{\sigma}^2_{Y,\text{red}})$.

To apply the estimation procedure to the bootstrap sample, we performed the principal component analysis to the regenerated observations $\{(t_{ij}, W_{ij}^*); j=1, \cdots, m_i, i=1, \cdots, n\}$, so that our test procedure automatically takes into account the estimation error in FPCA. After FPCA, we apply both the full model and the reduced model to the bootstrap sample and calculate the log likelihood ratio. The bootstrap procedure was repeated 1, 000 times.

Step 3. Compare the log likelihood ratio calculated from the data to the bootstrap quantities, and let the p-value be the upper percentile of the true log likelihood ratio among the bootstrap versions.

We applied this procedure to the colon carcinogenesis data, and the test yields a p-value of 0.044, which is indication that interaction does exist in these data. To see the effect of this interaction, we also present $\mathfrak{A}(t, \boldsymbol{\theta}^{\text{T}} \boldsymbol{Z}_1)$ in Figure 3. We can see, as the single-index value $\boldsymbol{\theta}^{\text{T}} \boldsymbol{Z}_1$ changes, the coefficient function for the functional predictor changes quite dramatically. Other inference procedures have been proposed in different semiparametric

models. Among them, Liang et al. (2009) proposed an empirical likelihood based inference for generalized partially linear models and Li and Liang (2008) proposed to use a semiparametric generalized likelihood ratio test to select significant variables in the nonparametric component in the generalized varying-coefficient partially linear models. How to adopt their testing ideas in our models warrants future research.

5.2.4 Further Investigation of Interaction—One interesting finding of our study is that when we apply the reduced model without interaction to the data, the only significant parameter is the main effect of diet. In other words, none of butyrate supplementation, proliferation and p27 show strong main effects. These factors only show their effects in the full model that includes an interaction.

To further confirm this conclusion, we also did another simulation. We bootstrapped crypts from the original data, but regenerated Y using our estimated full model with interactions. We then applied the reduced model to the simulated data sets. Again, none of the factors besides the fish oil diet showed significant main effects.

We now further look into the interactions between p27 and other covariates through the PC scores of p27 and $S(\cdot)$ values of the single index $\boldsymbol{\theta}^\mathrm{T}\boldsymbol{Z}_1$. To eliminate the influence of potential boundary effects, we focus on data points that are within the internal area of $S(\cdot)$. Precisely, we truncated 10% of the points from either side of the boundary of the range of $\boldsymbol{\theta}^\mathrm{T}\boldsymbol{Z}_1$. We then divided the data points into three subgroups: those that give high values of $S(\cdot)$ (S_high: $S > 1.5$); those ones give low values of $S(\cdot)$ (S_low: $S < -1.5$); and the ones that are in between (S_mid), where "high", "low", and "mid" imply positive, negative and zero values of S. In Figure 4, we plot the proportions of observations for all data values, the data in S_low and the data in S_high that receive fish oil diet and butyrate supplement, respectively. In Figure 5, we plot the three corresponding boxplots of proliferation values. We can see that the proportions of butyrate and fish oil for all data values and for those in S_high are similar, but for the S_low group, the proportion of butyrate supplements is higher while the proportion of fish oil is lower than the others. An equivalent statement applies to the distributions of proliferation. The S_low group tends to have higher proliferation while the proliferation of the S_high group is in general similar to the overall distribution.

To understand the interacting effects of p27 and $\boldsymbol{\theta}^\mathrm{T}\boldsymbol{Z}_1$ on the apoptotic levels through the p27 PC scores, we produce Figure 6. We first removed the estimated main effects by subtracting them from the apoptotic levels and refer to the adjusted values as apoptotic indices. We then dichotomized each of the two p27 PC scores according whether they belong to the top or bottom 50% of the scores; this practice allows us to produce 4 groups for the 2 PC-scores: PC1-Low, PC1-High, PC2-Low and PC2-High. An entry that belongs to PC1-Low tends to have a lower than average p27 and that in PC2-Low tends to have its p27 values lower in the top of the crypts. The equivalent implications apply to the PC1-High and PC2-High groups. We calculated the average apoptotic indices of these four PC groups against S_low, S_mid and S_high and plot them in Figure 6. The solid circle and square indicate the membership of PC1 and PC2, while the solid, dashed, dash-dotted and dotted lines indicate the membership of PC1-High, PC1-Low, PC2-High and PC2-Low, respectively. Figure 6 clearly shows that there are interactions between PC scores of p27 and $S(\boldsymbol{\theta}^\mathrm{T}\boldsymbol{Z}_1)$. When the p27 average is high (PC1-High) or it is higher on the top (PC2-High) of the crypt, whether there is a reduction of the apoptotic indices seems to be partly determined by whether there is a high proliferation level under butyrate supplementation. When the p27 average is low (PC1-Low), other covariates seem to have little effect on the change of the apoptotic index. On the other hand, if p27 levels are higher on the bottom of the crypt (PC2-Low), the high proliferation level under butyrate supplementation seems to induce an increment of the apoptotic index.

6 Conclusions

We have introduced a new class of functional generalized linear models that includes multivariate covariates and their interaction with the functional predictor. To achieve parsimony, the interaction is modeled semiparametrically with a single-index structure.

When the functional features related to the response variable are known basis functions, we proposed a two step procedure based on local estimation equations. We showed theoretically that the first step of the procedure produces a consistent estimator, while the second step estimator is more efficient and is asymptotically normal with a root n rate.

We further investigated the situation that the functional features are data-driven, i.e. the principal components of the process. We applied the functional principal component analysis of Yao et al. (2005) and Hall et al. (2006) to the discrete observations on the functional data, and plugged the estimated principal component scores into the two step procedure. We showed that when we have a large number of subjects, and enough number of observations on each curve, the initial estimator in our first step procedure is still consistent, while the second step estimator is still root-n consistent and asymptotically normal.

One important finding is that asymptotic variance of our second step estimator is larger than in the case that the functional features ξ are known. This extra variation in the estimator is due to the estimation error of the FPCA.

Our theoretical investigation is innovative since almost all the literature on functional data analysis that we are aware of tends to ignore the estimation error in FPCA. The only exception is perhaps the recent paper by Li, Li, Wang and Wang (2008), where estimation error in FPCA was taken into account, but their model did not include interactions between the functional predictor and other covariates. Our results verify that one should be aware of the extra variability when estimating FPCA scores.

Our methods were illustrated by a simulation study and applied to a colon carcinogenesis data set. In studying this FGLM model, we detected a previously unknown interaction of the response, p27, with the environmental covariates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Baladandayuthapani V, Mallick B, Hong M, Lupton J, Turner N, Carroll RJ. Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. Biometrics 2008;64:64–73. [PubMed: 17608780]

Cai T, Hall P. Prediction in functional linear regression. Annals of Statistics 2006;34:2159–2179.

Cardot H, Sarda P. Estimation in generalized linear models for functional data via penalized likelihood. Journal of Multivariate Analysis 2005;92:24–41.

Carroll RJ, Fan J, Gijbels I, Wand MP. Generalized partially linear single-index models. Journal of the American Statistical Association 1997;92(438):477–489.

Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement Error in Nonlinear Models: A Model Perspective. 2nd Edition. New York: Chapman & Hall; 2006.

Capra WB, Müller HG. An accelerated-time model for response curves. Journal of the American Statistical Association 1997;92:72–83.

Crainiceanu, CM.; Staicu, A.; Di, C. Generalized Multilevel Functional Regression. Johns Hopkins University, Dept. of Biostatistics Working Papers; 2008. Working Paper 173. http://www.bepress.com/jhubiostat/paper173.

Di C, Crainiceanu CM, Caffo BS, Punjabi NM. Multilevel functional principal component analysis. Annals of Applied Statistics. 2009 to appear.

Ferraty, F.; Vieu, P. Nonparametric Functional Data Analysis: Theory and Practice. New York: Springer-Verlag; 2006.

Hall P, Hosseini-Nasab M. On properties of functional principal components analysis. Journal of the Royal Statistical Society, Series B 2006;68:109–126.

Hall P, Müuller H-G, Wang J-L. Properties of principal component methods for functional and longitudinal data analysis. Annals of Statistics 2006;34:1493–1517.

Li E, Li Y, Wang N-Y, Wang N. Functional latent feature models for data with longitudinal covariate processes. 2008 Preprint.

Li KC. Sliced inverse regression for dimension reduction (with Discussion). Journal of the American Statistical Association 1991;86 316342.

Li R, Liang H. Variable selection in semiparametric regression modeling. Annals of Statistics 2008;36:261–286. [PubMed: 19122808]

Li Y, Wang N, Hong M, Turner N, Lupton J, Carroll RJ. Nonparametric estimation of correlation functions in spatial and longitudinal data, with application to colon carcinogenesis experiments. Annals of Statistics 2007;35:1608–1643.

Liang H, Qin YS, Zhang XY, Ruppert D. Empirical-likelihood-based inferences for generalized partially linear models. Scandinavian Journal of Statistics 2009;36:433–443.

Mack YP, Silverman BW. Weak and Strong Uniform consistency of kernel regression estimates, *Z. Wahrscheinlichkeitstheorie verw*. Gebiete 1982;61:405–415.

McCullagh, P.; Nelder, JA. Generalized Linear Models. 2nd Edition. New York: Chapman & Hall; 1989.

Morris JS, Wang N, Lupton JR, Chapkin RS, Turner ND, Hong MY, Carroll RJ. Parametric and nonparametric methods for understanding the relationship between carcinogen-induced DNA adduct levels in distal and proximal regions of the colon. Journal of the American Statistical Association 2001;96(455):816–826.

Morris JS, Vannucci M, Brown PJ, Carroll RJ. Wavelet-based non-parametric modeling of hierarchical functions in colon carcinogenesis. Journal of the American Statistical Association 2003;98(463): 573–583.

Müller H-G, Stadtmüller U. Generalized functional linear models. Annals of Statistics 2005;33:774–805.

Ramsay, JO.; Silverman, BW. Functional Data Analysis. 2nd Edition. New York: Springer-Verlag; 2005.

Rice J, Silverman B. Estimating the mean and covariance structure nonparametrically when the data are curves. Journal of the Royal Statistical Society, Series B 1991;53:233–243.

Xia Y, Tong H, Li WK, Zhu L. An adaptive estimation of dimension reduction space (with discussion). Journal of the Royal Statistical Society, Series B 2002;64:363–410.

Xia Y, Härdle W. Semi-parametric estimation of partially linear single-index models. Journal of Multivariate Analysis 2006;97:1162–1184.

Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association 2005;100:577–590.

Zhang J-T, Chen J. Statistical inferences for functional data. Annals of Statistics 2007;35:1052–1079.

Zhou L, Huang JZ, Carroll RJ. Joint modeling of paired sparse functional data using principal components. Biometrika 2008;95(3):601–619. [PubMed: 19396364]

Appendix: Asymptotic Covariance Matrices in Theorems 2 and 4

Throughout, $\eta\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, S(\cdot)\} = \eta\{\boldsymbol{\Theta}, S(\cdot)\}$ denotes the expression $\alpha_1^{\mathrm{T}}\boldsymbol{\xi} + S(\theta^{\mathrm{T}}Z_1)\alpha_2^{\mathrm{T}}\boldsymbol{\xi} + \beta^{\mathrm{T}}Z$, and the subindices "$i$" and "$ij$" of η indicate the replacement of each random variable by the

corresponding observations. Provided that the clarity of the presentation is preserved, we also leave out arguments of functions to simplify certain equations.

Define $\varepsilon_i = q_1\{\eta_i, Y_i\}$,

$$\mathscr{V}_1(\mathbf{z};\theta) = \mathrm{E}\{\rho_2(\eta_i)(\alpha_{2,0}^\mathrm{T}\xi_i)^2 | \theta^\mathrm{T} Z_{i1} = \theta^\mathrm{T}\mathbf{z}\},$$

$$\pi_{i1} = \rho_2(\eta_i)(\alpha_{2,0}^\mathrm{T}\xi_i)\left\{ \begin{array}{c} \xi_i \\ S(\theta_0^\mathrm{T} Z_{i1})\xi_i \\ Z_i \end{array} \right\}, \pi_{i2} = \rho_2(\eta_i)(\alpha_{2,0}^\mathrm{T}\xi_i)^2 Z_{i1},$$

$$\widehat{\pi}_1(\mathbf{z};\theta) = \mathrm{E}(\pi_{i1} | \theta^\mathrm{T} Z_{i1} = \theta^\mathrm{T}\mathbf{z}), \widehat{\pi}_2(\mathbf{z};\theta) = \mathrm{E}(\pi_{i2} | \theta^\mathrm{T} Z_{i1} = \theta^\mathrm{T}\mathbf{z}),$$

and $\mathscr{V}_3(\mathbf{z};\boldsymbol{\theta})$ is a symmetric matrix whose $(\ell_1, \ell_2)^{\text{th}}$ block is denoted by $\mathscr{V}_3^{[\ell_1,\ell_2]}$ for $\ell_1, \ell_2 = 1, 2$, and

$$\mathscr{V}_3^{[1,1]}(\mathbf{z};\theta) = \mathrm{E}\left[\rho_2(\eta_i)\left\{ S \begin{array}{c} \xi_i \\ (\theta_0^\mathrm{T} Z_{i1}) \xi_i \\ Z_i \end{array} \right\}^{\otimes 2} \Big| \theta^\mathrm{T} Z_{i1} = \theta^\mathrm{T}\mathbf{z}\right],$$

$$\mathscr{V}_3^{[1,2]}(\mathbf{z};\theta) = S^{(1)}(\theta_0^\mathrm{T}\mathbf{z})\{\mathrm{E}(\pi_{i1} Z_{i1}^\mathrm{T} | \theta^\mathrm{T} Z_{i1} = \theta^\mathrm{T}\mathbf{z}) - \widehat{\pi}_1(\mathbf{z};\theta)\mathbf{z}^\mathrm{T}\},$$

$$\mathscr{V}_3^{[2,1]}(\mathbf{z};\theta) = \{\mathscr{V}_3^{[1,2]}(\mathbf{z};\theta)\}^\mathrm{T},$$

$$\mathscr{V}_3^{[2,2]}(\mathbf{z};\theta) = \{S^{(1)}(\theta_0^\mathrm{T}\mathbf{z})\}^2 \left[\mathrm{E}\{\rho_2(\eta_i)(\alpha_{2,0}^\mathrm{T}\xi_i)^2 Z_{i1} Z_{i1}^\mathrm{T} | \theta^\mathrm{T} Z_{i1} = \theta^\mathrm{T}\mathbf{z}\} - \widehat{\pi}_2(\mathbf{z};\theta)\mathbf{z}^\mathrm{T} - \mathbf{z}\widehat{\pi}_2^\mathrm{T}(\mathbf{z};\theta) + \mathscr{V}_1(\mathbf{z};\theta)\mathbf{z}\mathbf{z}^\mathrm{T}\right].$$

Let

$$\mathscr{N}_n = n^{-1}\sum_{j=1}^n \varepsilon_i \left[\begin{array}{c} \xi_i - \mathscr{V}_1^{-1}(Z_{i1};\theta_0)(\alpha_{2,0}^\mathrm{T}\xi_i)\mathrm{E}\{\rho_2(\eta)(\alpha_{2,0}^\mathrm{T}\xi)\xi | \theta_0^\mathrm{T} Z_1 = \theta_0^\mathrm{T} Z_{i1}\} \\ S(\theta_0^\mathrm{T} Z_{i1})\xi_i - \mathscr{V}_1^{-1}(Z_{i1};\theta_0)(\alpha_{2,0}^\mathrm{T}\xi_i)\mathrm{E}\{\rho_2(\eta)(\alpha_{2,0}^\mathrm{T}\xi)S (\theta_0^\mathrm{T} Z_1)\xi | \theta_0^\mathrm{T} Z_1 = \theta_0^\mathrm{T} Z_{i1}\} \\ Z_i - \mathscr{V}_1^{-1}(Z_{i1};\theta_0)(\alpha_{2,0}^\mathrm{T}\xi_i)\mathrm{E}\{\rho_2(\eta)(\alpha_{2,0}^\mathrm{T}\xi)Z | \theta_0^\mathrm{T} Z_1 = \theta_0^\mathrm{T} Z_{i1}\} \\ (\alpha_{2,0}^\mathrm{T}\xi_i)S^{(1)}(\theta_0^\mathrm{T} Z_{i1})\{Z_{i1} - \mathscr{V}_1^{-1}(Z_{i1;\theta_0})\widehat{\pi}_2(Z_{i1};\theta_0)\} \end{array} \right],$$

and $\mathscr{N}_{1,n} = n^{-1}\sum_{i=1}^n \varepsilon_i(\widehat{\xi}_i - \xi_i)$ where

$$\varepsilon_i = \left[\begin{array}{c} \xi_i - \mathscr{V}_1^{-1}(Z_{i1};\theta_0)(\alpha_{2,0}^\mathrm{T}\xi_i)\mathrm{E}\{\rho_2(\eta)(\alpha_{2,0}^\mathrm{T}\xi)\xi | \theta_0^\mathrm{T} Z_1 = \theta_0^\mathrm{T} Z_{i1}\} \\ S(\theta_0^\mathrm{T} Z_{i1})\xi_i - \mathscr{V}_1^{-1}(Z_{i1};\theta_0)(\alpha_{2,0}^\mathrm{T}\xi_i)\mathrm{E}\left\{\rho_2(\eta)(\alpha_{2,0}^\mathrm{T}\xi)S(\theta_0^\mathrm{T} Z_1)\xi | \theta_0^\mathrm{T} Z_1 = \theta_0^\mathrm{T} Z_{i1}\right\} \\ Z_i - \mathscr{V}_1^{-1}(Z_{i1};\theta_0)(\alpha_{2,0}^\mathrm{T}\xi_i)\mathrm{E}\left\{\rho_2(\eta)(\alpha_{2,0}^\mathrm{T}\xi)Z | \theta_0^\mathrm{T} Z_1 = \theta_0^\mathrm{T} Z_{i1}\right\} \\ (\alpha_{2,0}^\mathrm{T}\xi_i)S^{(1)}(\theta_0^\mathrm{T} Z_{i1})\left\{Z_{i1} - \mathscr{V}_1^{-1}(Z_{i1;\theta_0})\widehat{\pi}_2(Z_{i1},\theta_0)\right\} \end{array} \right]$$
$$\times q_2(\eta_i, Y_i)\{\alpha_{1,0} + S_0(\theta_0^\mathrm{T} Z_{i1})\alpha_{2,0}\}^\mathrm{T}.$$

Finally, define $\mathscr{A}$, $\mathscr{B}$ and B_1 in Theorems 2 and 4 as:

$$\mathscr{A} = \mathrm{E}\{\mathscr{V}_3(Z_1;\theta_0)\}, \mathscr{B} = n \times \mathrm{cov}(\mathscr{N}_n), \mathscr{B}_1 = \lim_{n\to\infty} n \times \mathrm{cov}(\mathscr{N}_{1,n}).$$
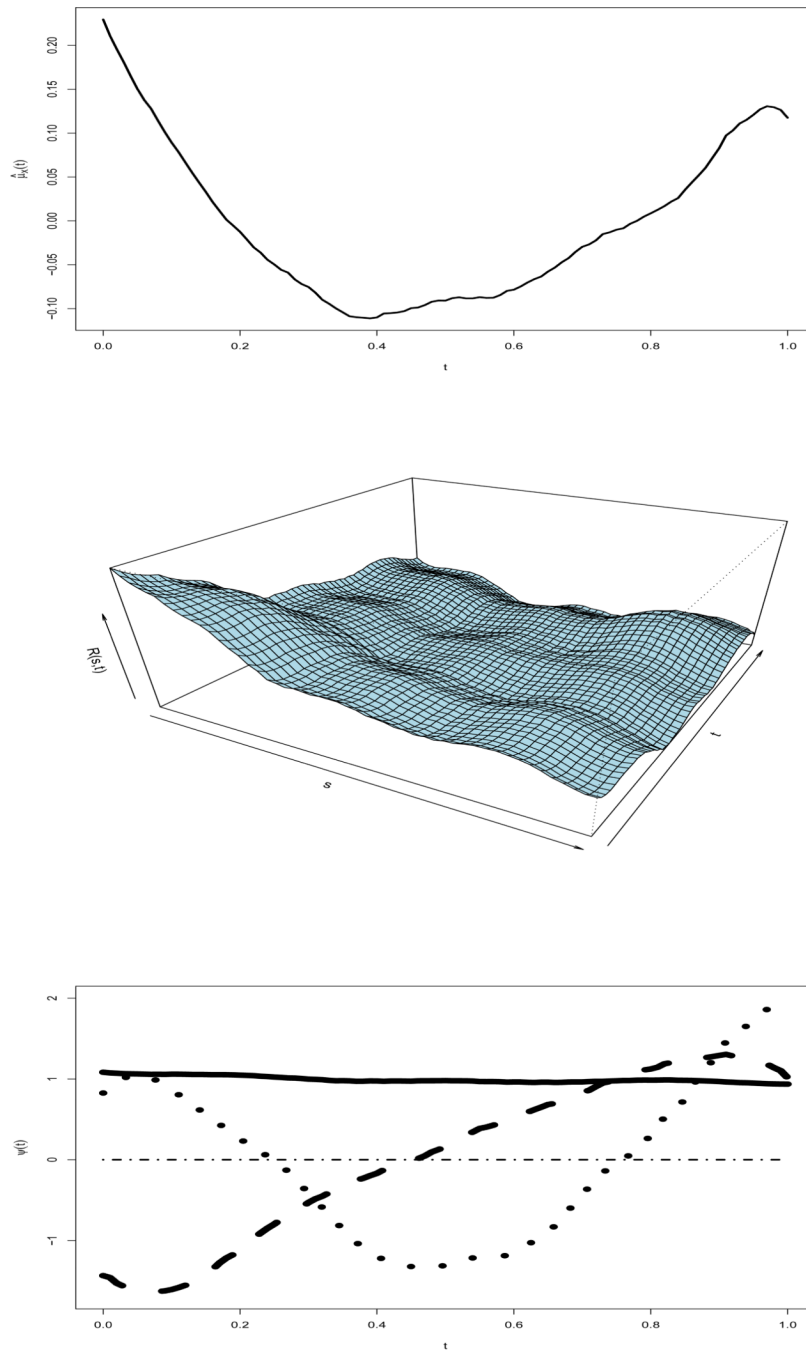
Figure 1.
Functional principal component analysis for the colon carcinogenesis p27 data. Top panel: local linear estimator of the mean curve $\mu_X(t)$; middle panel: local linear estimator of the covariance function; lower panel: the first 3 principal components. The solid curve, the dashed curve and the dotted curve are the first 3 principal components, respectively.
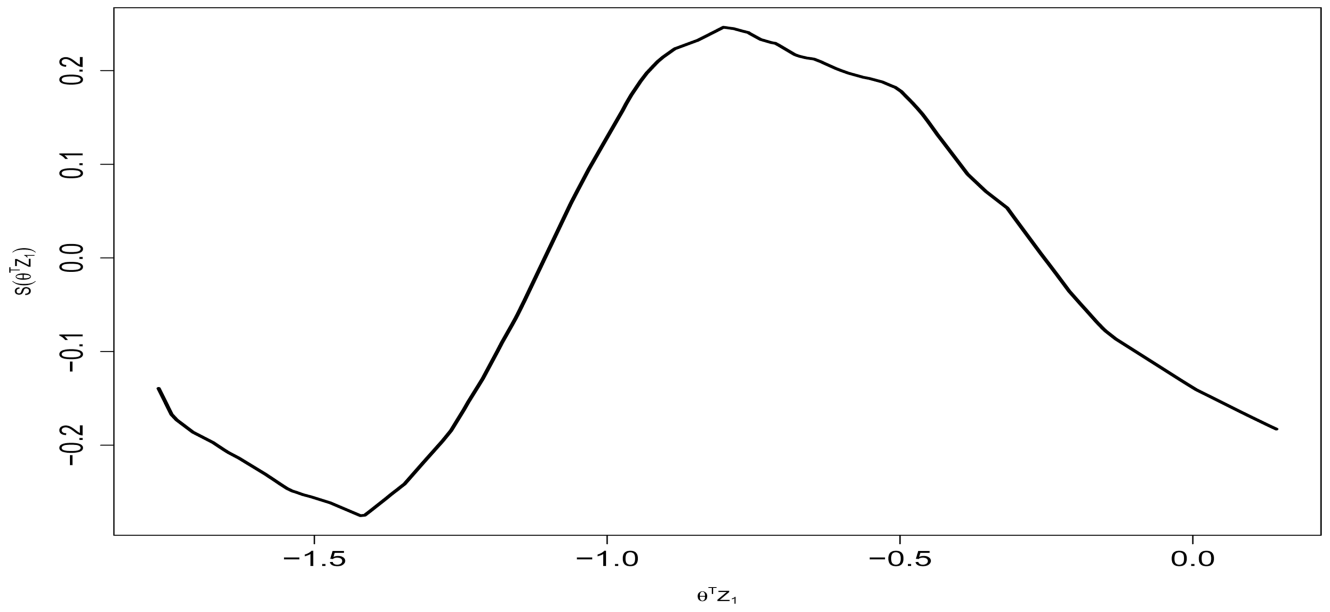
Figure 2.
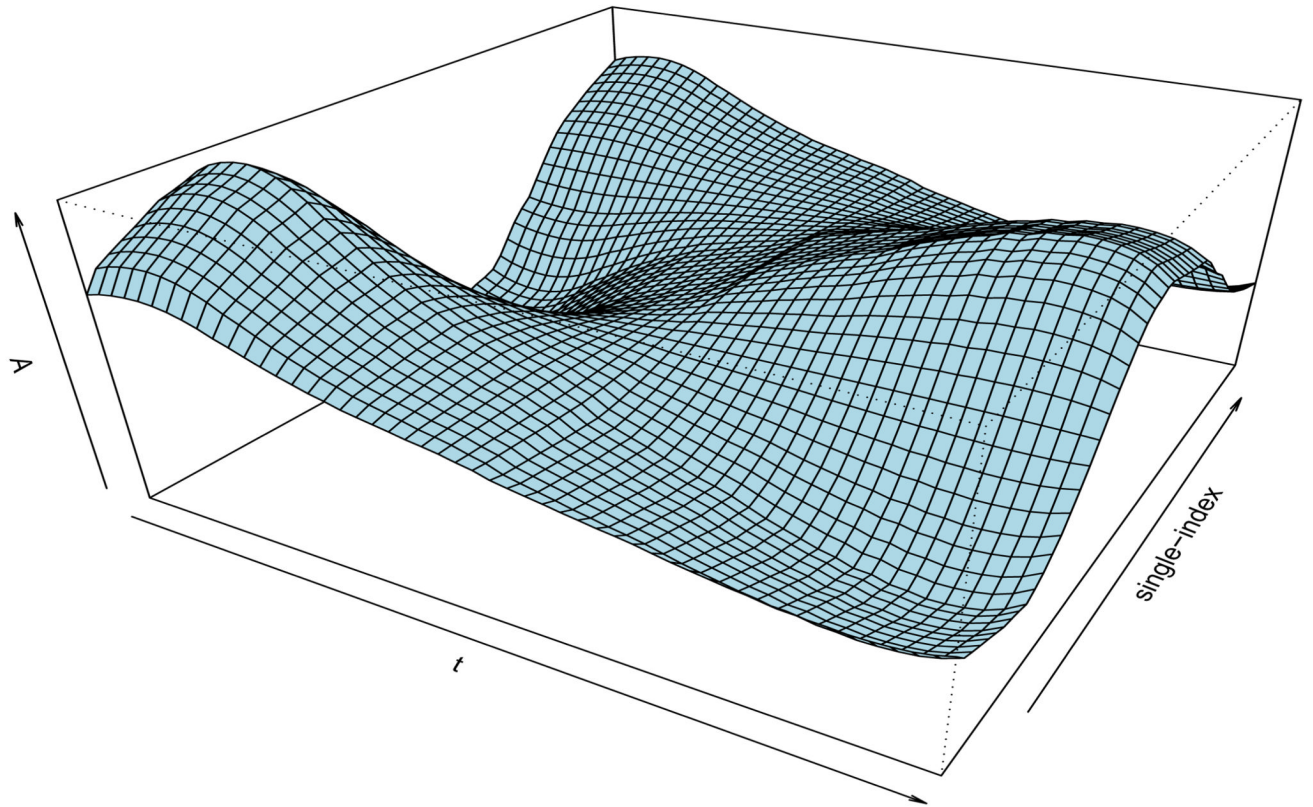Nonparametric estimator of $S(\cdot)$ in the colon carcinogenesis p27 data.

Figure 3.
Semiparametric estimator of $\mathfrak{A}(t, \boldsymbol{\theta}^{\mathrm{T}}\mathbf{Z}_1)$ in the colon carcinogenesis p27 data.
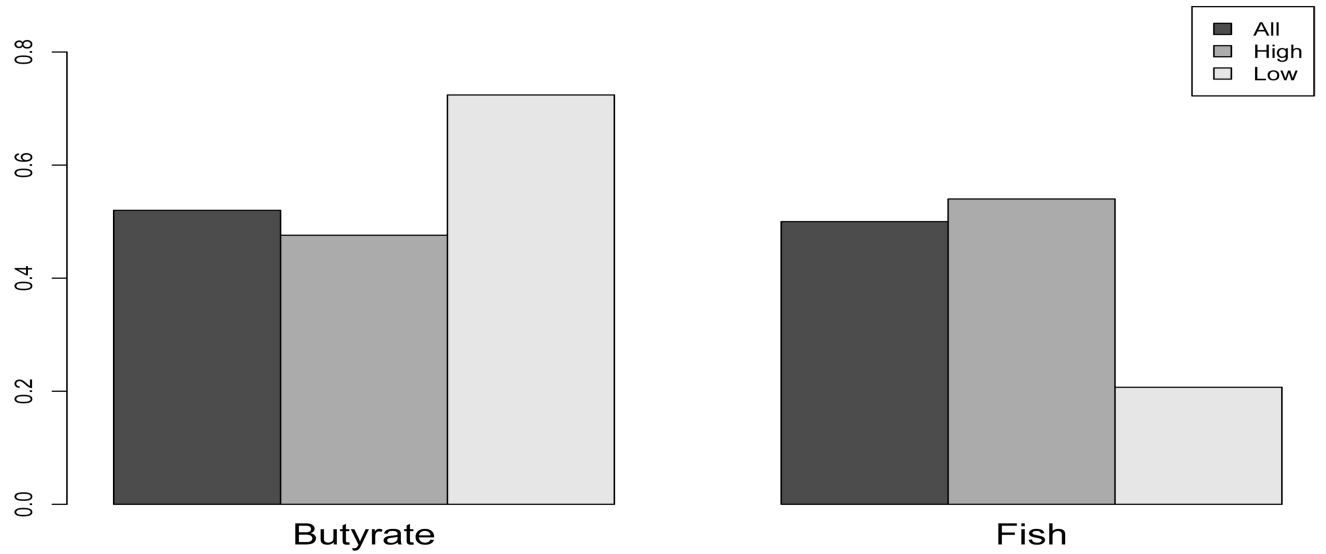
Figure 4.
Proportions of crypts receiving butyrate supplementation (left panel) and fish oil (right panel), depending on the levels of $S(\cdot)$. The left bar in each panel refers all crypts, the middle bar to the case that $S(\cdot)$ is high and the right to cases that $S(\cdot)$ is low.

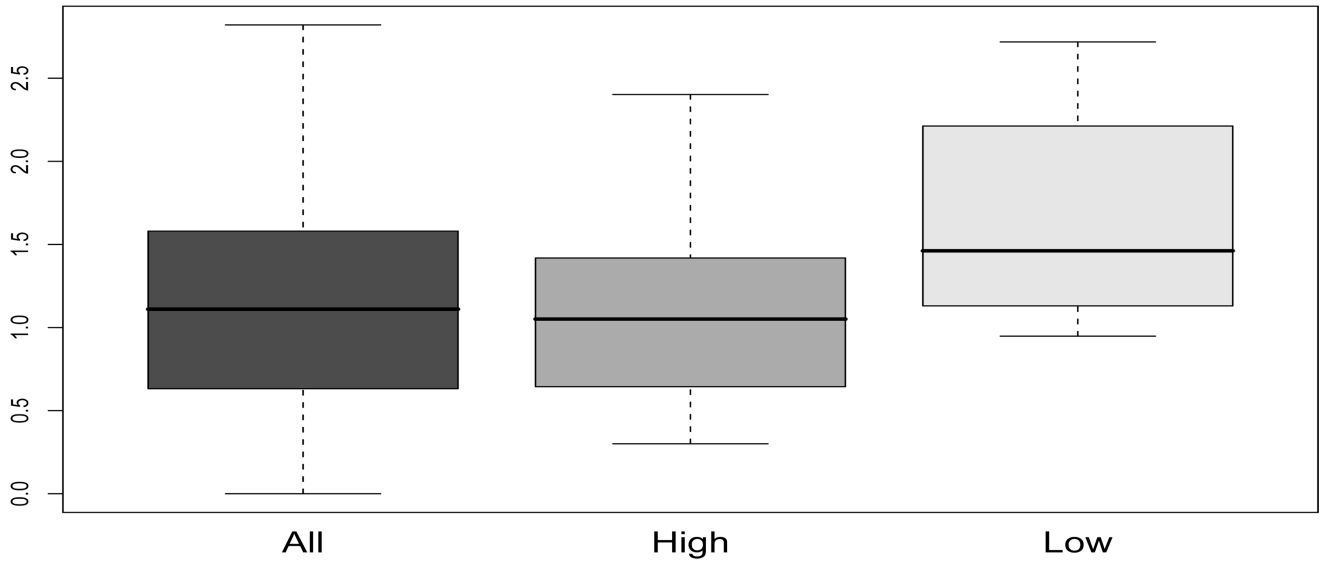Boxplot of Proliferation
for different S−values

Figure 5.
Boxplots of proliferation depending on values of $S(\cdot)$. Here "All", "High" and "Low" refer to all observations, those with high values of $S(\cdot)$ and those with low values of $S(\cdot)$.
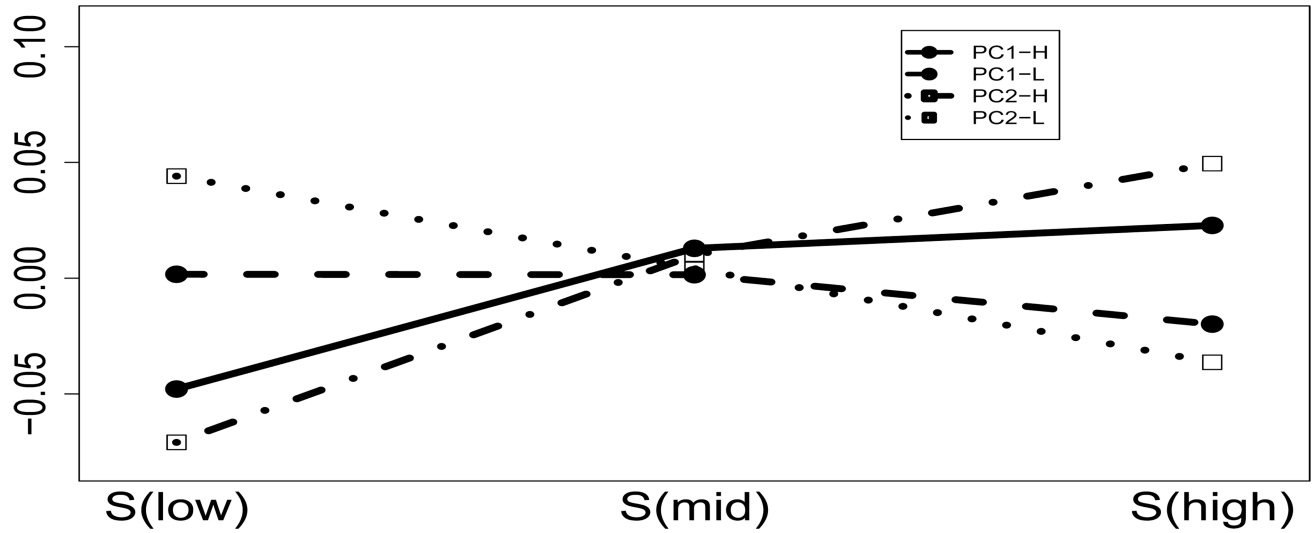
Figure 6.
Interaction between levels of the first two principal components and level of the function
$S(\cdot)$. Here "PC1" and "PC2" are the first and second principal components, respectively,
while subscripts "Low" and "High" indicate whether they are below or above the fiftieth
percentile. The terms S(low), S(mid) and S(high) refer to low, medium and high values of
$S(\cdot)$. The solid line refers to PC1-High, the dashed to PC1-Low, the dash-dotted to PC2-
High and the dotted to PC2-Low.

Table 1

Estimation results for the simulation study. The reduced model is misspecified because it does not include the interaction, and the biases observed in the reduced model are thus to be expected.

	Truth	Full Model				Reduced Model		
		MC Mean	MC SD	MC Bias		MC Mean	MC SD	MC Bias
β_0	−1	−1.0368	0.4181	−0.0368	β_0	−0.9381	0.3681	0.0619
β_1	2	2.0608	0.4925	0.0608	β_1	1.7558	0.4289	−0.2442
β_2	−2	−2.0197	0.4673	−0.0197	β_2	−1.7400	0.4057	0.2600
β_3	2	2.0744	0.2993	0.0744	β_3	1.8869	0.2705	−0.1131
α_{11}^*	1.7986	1.7525	0.1932	−0.0461	α_1	1.4958	0.1702	−0.3028
α_{12}^*	−0.0014	−0.0927	0.2009	−0.0913	α_1	−0.1626	0.1796	−0.1612
α_{21}	0.7071	0.6861	0.1887	−0.0210				
α_{22}	0.7071	0.6566	0.2511	−0.0505				
θ_1	0.7071	0.7012	0.1171	−0.0059				
θ_2	0.7071	0.6935	0.1177	−0.0136				

*
Because of the identifiability issue with α_1, the version of α_1 that the algorithm estimates is $\alpha_1^* = \alpha_1 + \mathrm{E}\{S(\theta^T Z_1)\}\alpha_2$.

Table 2

Variance inflation factors for using the true FPCA scores versus estimating them in the simulation study of Section 5. Displayed are the ratio of the variance for each parameter when estimating the FPCA scores to the variance when the true FPCA scores are known.

Parameter	Variance Inflation Factor
β_0	1.08
β_1	1.01
β_2	1.01
β_3	1.00
α_{11}	1.00
α_{12}	1.40
α_{21}	1.10
α_{22}	1.63
θ_1	1.04
θ_2	1.04

Table 3

Estimated parameters in the colon carcinogenesis data. The standard errors were estimated by bootstrap, and the p-values are for the null hypothesis that the parameter = 0.0. Here the two principal component values have $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12})^T$ and $\boldsymbol{\alpha}_2 = (\alpha_{21}, \alpha_{22})^T$. The subscripts "fish", "buty" and "prolif" refer to effects of a fish oil diet, butyrate supplementation and mean cell proliferation, respectively.

	α_{11}	α_{12}	α_{21}	α_{22}
Estimate	−0.0004	0.0235	−0.0480	0.9988
SE	0.0108	0.1003	0.1653	0.0767
p-value	0.9719	0.8145	0.7714	0.0000
	β_0	β_{fish}	β_{buty}	β_{prolif}
Estimate	0.2627	0.0514	0.0223	−0.0062
SE	0.0247	0.0193	0.0201	0.0135
p-value	0.0000	0.0078	0.2667	0.6484
	θ_{fish}	θ_{buty}	θ_{prolif}	
Estimate	0.4208	−0.7143	−0.5592	
SE	0.2847	0.2419	0.2005	
p-value	0.1394	0.0031	0.0053	