Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

SINGLE AND MULTIPLE INDEX FUNCTIONAL REGRESSION MODELS WITH NONPARAMETRIC LINK

By Dong Chen[‡], Peter Hall[*,‡,§] and Hans-Georg Müller[†,‡]

University of California, Davis[‡] *and University of Melbourne*[§]

Fully nonparametric methods for regression from functional data have poor accuracy from a statistical viewpoint, reflecting the fact that their convergence rates are slower than nonparametric rates for the estimation of high-dimensional functions. This difficulty has led to an emphasis on the so-called functional linear model, which is much more flexible than common linear models in finite dimension, but nevertheless imposes structural constraints on the relationship between predictors and responses. Recent advances have extended the linear approach by using it in conjunction with link functions, and by considering multiple indices, but the flexibility of this technique is still limited. For example, the link may be modelled parametrically or on a grid only, or may be constrained by an assumption such as monotonicity; multiple indices have been modeled by making finite-dimensional assumptions. In this paper we introduce a new technique for estimating the link function nonparametrically, and we suggest an approach to multi-index modeling using adaptively defined linear projections of functional data. We show that our methods enable prediction with polynomial convergence rates. The finite sample performance of our methods is studied in simulations, and is illustrated by an application to a functional regression problem.

1. Introduction. When explanatory variables are functions, rather than vectors, the problems of nonparametric regression and prediction are intrinsically difficult from a statistical viewpoint. In particular, convergence rates can be slower than the inverse of any polynomial in sample size, and so relatively large samples may be needed in order to ensure adequate performance. Fully nonparametric methods have been studied recently in functional-data regression and related problems (see, for example, [15], [13] and [21]). The slow convergence rates associated with these unstructured nonparametric approaches provide motivation for seeking nonparametric approaches that exploit a greater amount of structure in the data and are therefore expected to have better properties from a statistical perspective.

[*]Supported in part by an Australian Research Council Fellowship
[†]Supported in part by National Science Foundation grant DMS08-06199
AMS 2000 subject classifications: Primary 62G05, 62G08
Keywords and phrases: Functional Data Analysis, Generalized Functional Linear Model, Prediction, Smoothing

Advances in this direction include those made in [1], [8], [9], [17] and [20], where both parametric and nonparametric link functions were introduced in order to connect the response to a linear functional model in the explanatory variables. However, the flexibility of available link-function models is still rather limited. For example, although nonparametric link functions were considered in [20], the approaches considered there are restricted by the assumption of monotonicity, where the corresponding "Generalized Functional Linear Model" approach is based on a semiparametric quasi-likelihood based estimating equation, which includes known or unknown link and variance functions. In contrast, we are aiming here at models with one or several nonparametric link functions, ignoring possible heteroscedasticity of the errors. Our approach provides an alternative to the related methods in [2], where single-index functional regression models with general nonparametric link functions are considered that may be chosen non-monotonically and without shape constraints. The main differences are that our methodology includes the multi-index case, does not anchor the true parameter on a prespecified sieve, and that we provide a detailed theoretical analysis of a direct kernel-based estimation scheme that culminates in a convergence result that establishes a polynomial rate of convergence.

Beyond demonstrating that our approach enables prediction with polynomial accuracy, we also include generalizations to iteratively fitted multiple index models, founded on a sequence of linear regressions. Here we borrow ideas from dimension reduction in models that involve high-dimensional, but not functional, data. When the link function is nonparametric, the intercept term in functional linear regression loses its relevance because it is incorporated into the link. The slope function is still potentially of interest, but the viewpoint taken in this paper is predominately one of prediction rather than slope estimation, and in particular our theory focuses directly on the prediction problem. We refer to the papers by [4], [5] and [7] for further discussion of these objectives in the context of the functional linear model.

We introduce our model and estimation methodology in Section 2. Theoretical results regarding the polynomial convergence rate are discussed in Section 3, while algorithmic details are described in Section 4, which also includes an illustration of the proposed methods with an application to spectral data. Simulation results are reported in Section 5. Detailed assumptions and proofs can be found in the Appendix.

2. Model and Methodology.

2.1. *Model.* Suppose we observe data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, independent and identically distributed as (X, Y), where X is a random function

defined on a compact interval $\mathcal{I}$ and Y is a scalar. We anticipate that (X, Y) is generated as

$$Y = g(X) + \text{error}\,, \tag{2.1}$$

where g is a smooth functional and the error has zero mean, finite variance and is uncorrelated with X. The model at (2.1) admits many interpretations and generalizations, where (for instance) X is a multivariate rather than univariate function. For example, X might be (Z, Z'), where Z is a univariate function and Z' its derivative. To simplify the developments, we shall focus on problems where X is a univariate function of a single variable. Models and methodology in more general settings are readily developed from the single-variable case. Our approach is described in detail for situations where the trajectories of functional predictors can be assumed to be fully observed, for example due to smoothness such as for the Tecator data which we analyze with the proposed methods in Section 4.2; it can be extended to cases with densely and regularly measured trajectories, where measurements may be subject to i.i.d. noise with finite fourth order moments. This extension requires sufficiently dense measurement designs, such that smoothness assumptions coupled with suitable smoothing methods lead to sufficiently fast uniform rates of convergence when pre-smoothing the data to generate smooth trajectories. Such an extension will not be feasible for functional data for which only sparse and noisy measurements are available.

The case where g, in (2.1), is a general functional, even a very smooth functional, can have serious drawbacks from the viewpoint of practical function estimation, since the problem of estimating such a g is inherently difficult from a statistical viewpoint. In particular, convergence rates of estimators in this case are generally slower than the inverse of any polynomial in sample size. Therefore, unless the data set is very large, it can be particularly difficult to estimate g effectively. In this respect the commonly assumed functional linear model, where $g(x) = \alpha + \int_{\mathcal{I}} \beta\, x$, α is a scalar and β is a regression parameter function, offers substantial advantages, for example polynomial convergence rates and even, on occasion, root-n consistency. However, the linear-model assumption is often too restrictive in practical applications.

An alternative approach is to place the linear model inside a link function, for example defining

$$g(x) = g_1\biggl(\alpha + \int_{\mathcal{I}} \beta\, x\biggr)\,, \tag{2.2}$$

although this too is restrictive unless we select the link in a very adaptive manner. We suggest choosing the link function g_1 nonparametrically. In this

case the intercept parameter, α, in (2.2) is superfluous; it can be replaced by zero, and its effect incorporated into g_1. Therefore we actually fit the model

$$g(x) = g_1\Big(\int_{\mathcal{I}} \beta\, x\Big), \tag{2.3}$$

where g_1 is subject only to smoothness conditions, and to ensure identifiability, we require a condition on the "scale" of β, which we choose as $\int_{\mathcal{I}} \beta^2 = 1$. The sign of β can be determined arbitrarily.

2.2. *Methodology.* We estimate the parameter function β and the link function g_1 in the model at (2.3), using least-squares in conjunction with local-constant or local-linear smoothing as follows. To obtain g_1, we will use a scatterplot smoother which we implement as local constant or local linear weighted least squares smoothing. Given a parameter function β, the scatterplots targeting the nonparametric regression $g_1(z) = E(Y | \int_{\mathcal{I}} \beta\, X = z)$, consists of the data pairs $(\int_{\mathcal{I}} \beta\, X_i, Y_i)_{i=1,\ldots,n}$. Omitting the predictor X_j when predicting the response at $\int_{\mathcal{I}} \beta\, X_j$, averaging least squares smoothers constructed for predicting at the observed predictor levels X_j are then obtained by choosing intercept parameters ζ_j and slope parameters ϑ_j to minimize

$$\sum_{i,j\,:\,i\neq j} (Y_i - \zeta_j)^2 K\Big\{h^{-1} \int_{\mathcal{I}} \beta\,(X_i - X_j)\Big\} \quad \text{or} \tag{2.4}$$

$$\sum_{i,j\,:\,i\neq j} \Big\{Y_i - \Big(\zeta_j + \vartheta_j \int_{\mathcal{I}} \beta\, X_i\Big)\Big\}^2 K\Big\{h^{-1} \int_{\mathcal{I}} \beta\,(X_i - X_j)\Big\},$$

in the local-constant and local-linear cases, respectively, where K is a kernel function and h is a bandwidth.

Defining $K_{ij} = K\{h^{-1} \int_{\mathcal{I}} \beta\,(X_i - X_j)\}$, $\bar{X}_j = (\sum_{i\,:\,i\neq j} X_i\, K_{ij}) / \sum_{i\,:\,i\neq j} K_{ij}$ and $\bar{Y}_j = (\sum_{i\,:\,i\neq j} Y_i K_{ij}) / \sum_{i\,:\,i\neq j} K_{ij}$, one finds that the minimia of (2.4), for any given β, are

$$\sum_{i,j\,:\,i\neq j} (Y_i - \bar{Y}_j)^2\, K_{ij} \quad \text{or} \quad \sum_{i,j\,:\,i\neq j} \Big\{Y_i - \bar{Y}_j - \hat{\vartheta}_j \int_{\mathcal{I}} \beta\,(X_i - \bar{X}_j)\Big\}^2 K_{ij}. \tag{2.5}$$

The minimizers $\hat{\zeta}_j$ are given by $\hat{\zeta}_j = \hat{\zeta}_j(\beta) = \bar{Y}_j$ in the local-constant case and in the local-linear case minimizers $\hat{\zeta}_j$ and $\hat{\vartheta}_j$ are given by

$$\hat{\zeta}_j(\beta) = \bar{Y}_j - \hat{\vartheta}_j(\beta) \int_{\mathcal{I}} \beta \bar{X}_j, \tag{2.6}$$

$$\hat{\vartheta}_j(\beta) = \frac{\sum_{i\,:\,i\neq j} \{\int_{\mathcal{I}} \beta(X_i - \bar{X}_j)\}\,(Y_i - \bar{Y}_j)\, K_{ij}}{\sum_{i\,:\,i\neq j} \{\int_{\mathcal{I}} \beta\,(X_i - \bar{X}_j)\}^2\, K_{ij}}, \quad 1 \le j \le n.$$

Summarizing, the criteria at (2.4) are based on averaging local-constant and local-linear fits to $g_1(\int_{\mathcal{I}} \beta\, x)$ at $x = X_j$, averaging over X_j, where the respective fits are computed from the data $X_1, \ldots, X_n$, excluding X_j. The resulting approximations to $g_1(\int_{\mathcal{I}} \beta\, X_j)$, for a given β, are $\bar{Y}_j$ and $\bar{Y}_j + \hat{\vartheta}_j(\beta) \int_{\mathcal{I}} \beta\,(X_j - \bar{X}_j)$, respectively, with $\hat{\vartheta}_j(\beta)$ given by (2.6).

It remains to specify our final estimates. We estimate β by conventional least-squares, aiming to minimize the sum of squared differences between Y_j and the just-mentioned approximations:

$$S(\beta) = \sum_{j=1}^{n} (Y_j - \bar{Y}_j(\beta))^2 \quad \text{or} \tag{2.7}$$

$$S(\beta) = \sum_{j=1}^{n} \Big\{ Y_j - \bar{Y}_j(\beta) - \hat{\vartheta}_j(\beta) \int_{\mathcal{I}} \beta\,(X_j - \bar{X}_j) \Big\}^2,$$

subject to $\int_{\mathcal{I}} \beta^2 = 1$ and with $\hat{\vartheta}_j(\beta)$ as in (2.6). This problem is most conveniently solved by expanding $\beta = \sum_{1 \le k \le r} b_k\, \psi_k$, where $\psi_1, \psi_2, \ldots$ is an orthonormal basis and r denotes a frequency cut-off, choosing the generalized Fourier coefficients b_k to minimize $S(\beta)$. This gives estimators $\hat{b}_1, \ldots, \hat{b}_r$ of $b_1, \ldots, b_r$ respectively, and from those we may compute our estimator of β:

$$\hat{\beta} = \sum_{k=1}^{r} \hat{b}_k\, \psi_k, \quad \text{subject to} \quad \sum_{k=1}^{r} \hat{b}_k^2 = 1. \tag{2.8}$$

The basis $\psi_1, \psi_2, \ldots$ can be chosen as a fixed basis such as one of various orthonormal polynomial systems or the Fourier basis, or could be another sequence altogether, chosen for computational convenience. We note that it does not matter for the validity of our results whether the basis functions are fixed or random. Therefore the basis can be chosen as estimated eigenfunction basis, as long as the estimated eigenfunctions are orthonormal. We note that irrespective of how it is constructed, the selected basis needs to be such that condition (3.4) below is satisfied for the generalized Fourier coefficients of β, while the additionally needed conditions (3.5), (3.6) depend only on properties of β and X but not on the choice of the basis. The condition at (3.4) requires a polynomial decay rate (of arbitrary order) for the tail sums of the Fourier coefficients of β, which is slightly stronger than the convergence of the tail sums to 0 that is implied by the square integrability of β. Since we do not assume prior knowledge about β, no particular basis is preferrable in this regard a priori. In any case, the theory applies if (3.4) holds for the selected basis. In practice, one would choose a basis based on how well the representation of β works in typical applications. We found the

choice of estimated eigenfunctions for representing β particularly convenient for our applications and our implementation is therefore using this basis.

We note that the criteria at (2.4) are not directly comparable with those at (2.7), not least because in (2.4) we are fitting g_1 locally and in (2.7) we are fitting β globally. Reflecting these two different contexts, each residual squared error in both criteria in (2.4) has a local kernel weight, whereas each residual squared error in the criteria in (2.7) has a constant weight.

Having computed $\hat\beta$ we estimate the univariate function $g_1(u)$ by conventional local-constant or local-linear regression on the pairs $(\int_{\mathcal{I}} \hat\beta\, X_i, Y_i)$, for $1 \le i \le n$. In particular, in the local-constant case we take

$$\hat g_1(u) = \biggl\{ \sum_{i=1}^n Y_i\, K_i(u) \biggr\} \Big/ \biggl\{ \sum_{i=1}^n K_i(u) \biggr\}, \tag{2.9}$$

where $K_i(u) = K\{h^{-1}\,(\int_{\mathcal{I}} \hat\beta\, X_i - u)\}$; in the local linear setting we choose $\zeta = \hat\zeta$ and $\vartheta = \hat\vartheta$, both of which can also be viewed as functions of u, to minimize $\sum_i \{Y_i - (\zeta + \vartheta \int_{\mathcal{I}} \hat\beta\, X_i)\}^2\, K_i(u)$, and then put $\hat g_1(u) = \hat\zeta + \hat\vartheta\, u$. Several aspects of this algorithm can be modified to improve its performance. For example, noting that the ratio on the right-hand side of (2.6) will likely be unstable if the denominator is based on a relatively small number of terms, we might restrict the sum over j in either formula in (2.5) to values of that index for which $\sum_{i\,:\,i\ne j} K_{ij} \ge \lambda$, where $\lambda > 0$ denotes a sufficiently large threshold, and repeat this restriction in the case of (2.7). Problems caused by a too-small denominator can be especially serious in the case of functional data, since sample sizes there are typically relatively small.

If we take the view that the problem of interest is that of estimating g for the purpose of prediction, and that estimating β and g_1 in their own right is of relatively minor interest, then standard cross-validation can be used to choose simultaneously the smoothing parameters h and r. In section 3 we adopt the perspective of prediction, and show that in that context the estimator $\hat g$ approximates g at a rate that is polynomially fast as a function of sample size.

2.3. *Multiple index models.* The model at (2.3) can be generalized by taking g_1 to be a p-variate function

$$g(x) = g_1\biggl(\int_{\mathcal{I}} \beta_1\, x, \ldots, \int_{\mathcal{I}} \beta_p\, x \biggr), \quad \int_{\mathcal{I}} \beta_j^2 = 1 \quad \text{for} \quad 1 \le j \le p. \tag{2.10}$$

However, given the relatively small sample sizes often encountered in functional data analysis, focusing on the function at (2.10), with $p \ge 2$, will

often lead to estimators with high variability. An alternative, p-component functional multiple index model, such as

$$g(x) = g_1\Big(\int_{\mathcal{I}} \beta_1\, x\Big) + \ldots + g_p\Big(\int_{\mathcal{I}} \beta_p\, x\Big)\,, \tag{2.11}$$

is arguably more attractive. This class of models has been considered by [18], who referred to them as "Functional Adaptive Models". The approach of James and Silverman was restricted to the parametric case by requiring the functional predictors x_i as well as the link functions g_j to be elements of a finite-dimensional spline space, excluding nonparametric (infinite-dimensional) link and predictor functions. Such a fully parametric framework allows the use of a likelihood based approach to fitting these models, establishing identifiability by extending previous results for the vector case [6].

Since our main goal is prediction and not model identification, we are not primarily concerned with identifiability issues and do not emphasize specific identifiability conditions for the models we consider. The models at (2.10) and (2.11) in fact are not identifiable without further restrictions. To appreciate why, note that the order of the components on the right-hand side of (2.10), or of the functions on the right-hand side of (2.11), could be permuted without affecting the model. This problem does not arise for conventional multivariate or additive models, where the arguments of the functions are predetermined as the components of the explanatory variable x. While this difficulty can be overcome in a variety of ways, using a recursive additive model is attractive on both statistical and computational grounds. We now give background for that approach.

It is not uncommon in statistics to pragmatically alter a difficult problem to one that is simpler. Indeed, the introduction of additive models is typically motivated in that manner. Thus, we could generalize the problem of estimating a link function g, and a slope function β, in (2.1), subject only to smoothness conditions, to that of estimating the intrinsically simpler functions defined at (2.11). Alternatively, and more appropriately from the perspective of general inference, we would seek to estimate g in (2.10) not because we felt that those functions were identical to g in (2.1), but because they were relatively accessible approximations to g. Taking this view of the problem of estimating, or rather, approximating, the function g in (2.1), and accepting that the p-additive function at (2.11) is more likely to be practicable in functional data analysis than the p-variate function at (2.10), we suggest fitting the g in (2.11) recursively, for steadily increasing values of p. This "backfitting" approach borrows an idea from projection pursuit regression, to use recursive, low-dimensional, projection-based approximations.

In particular, taking $g_1^0 = g^0$ where g^0 denotes the true value of g at (2.1), we choose the function g_1 of a single variable, and the function β_1, to minimize, in the case $j = 1$, the expected value

$$E\Big\{g_j^0(X) - g_j\Big(\int_{\mathcal{I}} \beta_j\, X\Big)\Big\}^2, \quad \text{subject to} \quad \int_{\mathcal{I}} \beta_j^2 = 1. \tag{2.12}$$

More generally, if we have calculated β_{j-1} and g_{j-1}, and previously defined $g_{j-1}^0(x)$, then we may define g_j^0 by $g_j^0(x) = g_{j-1}^0(x) - g_{j-1}(\int_{\mathcal{I}} \beta_{j-1}\, X)$ and choose g_j and β_j to minimize the quantity at (2.12).

In the next section we shall show how to calculate estimators $\hat{g}_j$ and $\hat{\beta}_j$ of g_j and β_j, respectively, for $j \geq 1$. Note that we do not claim to consistently estimate g, in (2.1), unless that function has exactly the form at (2.3) (in which case our estimator is $\hat{g} = \hat{g}_1$, defined in section 2.2). Instead we suggest developing consistent estimators of successive approximations to $g(x)$, i.e. of

$$\begin{aligned} &g_1\Big(\int_{\mathcal{I}} \beta_1\, x\Big)\,,\ g_1\Big(\int_{\mathcal{I}} \beta_1\, x\Big) + g_2\Big(\int_{\mathcal{I}} \beta_2\, x\Big)\,, \\ &\qquad\qquad g_1\Big(\int_{\mathcal{I}} \beta_1\, x\Big) + g_2\Big(\int_{\mathcal{I}} \beta_2\, x\Big) + g_3\Big(\int_{\mathcal{I}} \beta_3\, x\Big)\,, \ldots\,. \end{aligned} \tag{2.13}$$

2.4. *Estimation in functional multiple index models.* Here we generalize the methodology in section 2.2 so that it permits estimation of the functions $g_1, g_2, \ldots$ in (2.12). Assume we are fitting a p–index model. The recursive fitting procedure means once we have estimators $\hat{\beta}_j$ and $\hat{g}_j$, for $1 \leq j \leq k-1 < p$, of the functions β_j and g_j defined in the paragraph containing (2.12), we take $Y_i(k) = Y_i - \hat{g}_1(X_i) - \ldots - \hat{g}_{k-1}(X_i)$, and use the methodology in section 2.2 but with $Y_i(k)$ replacing Y_i, obtaining an estimator $\hat{\beta}$, on this occasion actually an estimator $\hat{\beta}_k$ of β_k, and an estimator $\hat{g}$, which is really an estimator $\hat{g}_k$ of g_k. The quantity $\hat{g}_1(\int_{\mathcal{I}} \hat{\beta}_1\, x) + \cdots + \hat{g}_p(\int_{\mathcal{I}} \hat{\beta}_p\, x)$ is our estimate of the p-index model from the recursive fitting procedure.

A further refinement that leads to smaller prediction errors is backfitting, which uses the recursive fits described above as a starting point. Once these fits are obtained, further updates are obtained iteratively by revisiting and updating one index after another, presuming that the remaining $p-1$ indexes are fixed. The iterative updating of individual indices is itself iterated until indices change only little. This is implemented in a similar way as described in [6] for a traditional multiple index model with monotone link functions. Denoting the estimates obtained from the initial recursive fitting procedure by $\hat{g}_1^0(\int_{\mathcal{I}} \hat{\beta}_1^0\, x) + \ldots + \hat{g}_p^0(\int_{\mathcal{I}} \hat{\beta}_p^0\, x)$, then for the d-th iteration, iterating also

through the increasing sequence $k = 1, 2, \ldots, p$, one uses

$$Y_i^d(k) = Y_i - \sum_{j<k} \hat{g}_j^d\Big(\int_{\mathcal{I}} \hat{\beta}_j^d\, X_i\Big) - \sum_{j>k} \hat{g}_j^{d-1}\Big(\int_{\mathcal{I}} \hat{\beta}_j^{d-1}\, X_i\Big) \tag{2.14}$$

to replace Y_i for fitting $\hat{g}_k^d(\int_{\mathcal{I}} \hat{\beta}_k^d\, x)$. The iterative backfitting procedure is stopped once the relative differences between $\hat{\beta}_1^{d-1}$ and $\hat{\beta}_1^d$ fall below a prespecified threshold or a maximum number of iterations is reached.

3. Polynomial Convergence Rate. The main result in this section establishes that, if the linear model is linked to the response variable as in (2.3), if a Hölder smoothness condition on the link function g_1 is assumed, and if we ask of the generalized Fourier expansion $\beta = \sum_{k\geq 1} b_k\, \psi_k$ that it converges polynomially fast at a sufficiently rapid rate, then the predictor $\hat{g}$ converges to g at a polynomial rate. That property distinguishes the approach suggested in this paper from fully nonparametric methods that impose only smoothness conditions on the function g, in (2.1), but have much slower convergence rates for the predictor. We give explicit theory only in the local-constant case, since, as argued at the end of section 2.2, that approach is particularly appropriate when dealing with functional data. The local-linear setting can be treated similarly.

We assume that independent and identically distributed data pairs (X_i, Y_i) are generated by the model discussed in section 2:

(3.1) $Y_i = g(X_i) + \epsilon_i$, where the X_i's are square-integrable random functions supported on the compact interval $\mathcal{I}$, g is a real-valued functional given by $g(x) = g_1(\int_{\mathcal{I}} \beta^0\, x)$, g_1 is a real-valued function of a single variable, β^0 enjoys the property $\int_{\mathcal{I}} {\beta^0}^2 = 1$ and denotes the true value of the square-integrable function β, and the errors ϵ_i are independent of the X_i's and of one another, and have zero mean.

The only assumption we make of g_1 is that it is bounded and smooth:

(3.2) g_1 is bounded and satisfies a Hölder continuity condition: $|g_1(u) - g_1(v)| \leq D_1\, |u - v|^{a_1}$ for all u and v, where $a_1, D_1 > 0$.

The assumption that g_1 is bounded can be relaxed. For example, if the functions X_i are bounded with probability 1 then $\int_{\mathcal{I}} \beta^0 X_i$ is uniformly bounded, and so the distribution of the response variables Y_i depends only on the values that g_1 takes on a particular compact interval. We can extend g_1 from that interval to the whole real line in such a way that the extended version of g_1 is bounded and has a bounded derivative. More generally, if $\sup_{1 \le i \le n} \|X_i\|$ grows at rate $O(n^\eta)$, for all $\eta > 0$, where $\|X\|$ denotes the L_2 norm of X (for example, this condition holds if X is a Gaussian process), and if $\sup_{|x| \le u} |g_1(x)|$ grows at no faster than a polynomial rate as u diverges, then only minor modifications of our proof of the theorem are required to establish Theorem 1.

Let X have the common distribution of the random functions X_i in the model at (3.1). We ask that $\|X\|$ have at least a small, fractional moment, and that all moments of the error distribution be finite. In particular:

(3.3) $E(\|X\|^\eta) < \infty$ for some $\eta > 0$, and $E(|\epsilon|^m) \le (D_2\, m)^{a_2 m}$ for all integers $m \ge 1$, where a_2, D_2 denote positive constants.

The condition $E|\epsilon|^m \le (D_2\, m)^{a_2 m}$ is satisfied by distributions the tails of which decrease at rate at least $\exp(-C_1\, x^{C_2})$, for constants $C_1, C_2 > 0$, provided we choose $a_2 > 1/C_2$. In particular, the condition is satisfied by exponential and Gaussian distributions, and also, in the case $C_2 < 1$, by many distributions that do not have finite moment generating functions.

Write $f(\cdot \,|\, \beta)$ for the probability density of $\int_{\mathcal{I}} \beta X$. Given an orthonormal basis $\psi_1, \psi_2, \ldots$ for the class $L_2(\mathcal{I})$ of square-integrable functions on $\mathcal{I}$, express a general function $\beta \in L_2(\mathcal{I})$ with $\int_{\mathcal{I}} \beta^2 = 1$ as $\beta = \sum_{k \ge 1} b_k\, \psi_k$, where $\sum_{k \ge 1} b_k^2 = 1$. For constants $a_3, a_4, B, D_3, D_4, D_5 > 0$, we shall assume that:

$$\sum_{k=r+1}^{\infty} b_k^2 \le\ D_3\, (1+r)^{-B} \quad \text{for all } r \ge 1\,, \tag{3.4}$$

$$\sup_{\beta \in \mathcal{B}\,;\, x} f(x \,|\, \beta) < \infty\,, \tag{3.5}$$

$$\sup_{\beta \in \mathcal{B}} P\Big\{ f\big(\textstyle\int_{\mathcal{I}} \beta X - u \,\big|\, \beta\big) \le D_4\, \delta^{a_3} \quad \text{for all} \quad |u| \le \delta \Big\} \le D_5\, \delta^{a_4}\,, \tag{3.6}$$

where (3.6) holds for all sufficiently small $\delta > 0$. Condition (3.4) is standard; it asks that the generalized Fourier coefficients of β decay at least polynomially fast, in a weak sense. To appreciate the motivation for (3.5) and (3.6), observe that if X is a Gaussian process for which the covariance operator has eigenvalues $\theta_1 \ge \theta_2 \ge \ldots \ge 0$ and respective eigenfunctions $\phi_1, \phi_2, \ldots$, then $f(\cdot \,|\, \beta)$ is the $\mathrm{N}(a, \varsigma^2)$ density, where $a = \int_{\mathcal{I}} \beta\, E(X)$, $\varsigma^2 = \sum_{k \ge 1} \theta_k\, b_k^2$

and $b_k = \int_{\mathcal{I}} \beta\, \phi_k$. Then (3.6) is obtained by using well-known tail bounds for the Gaussian distribution function Φ with standard Gaussian density ϕ. It follows that (3.5) and (3.6) hold whenever $0 < a_4 \le a_3 < \infty$ and $\mathcal{B}$ is a class of functions β for which $\sum_{k \ge 1} \theta_k\, b_k^2$ is bounded away from zero and infinity, and for which $\sum_{k \ge 1} b_k^2 = 1$. Our use of the principal component basis in this example serves only to show the reasonableness of conditions (3.5) and (3.6), which of course do not depend on choice of basis. It does not imply that the basis $\psi_1, \psi_2, \ldots$ should be identical to $\phi_1, \phi_2, \ldots$.

Of the kernel K and bandwidth h we ask that:

(3.7) K is nonnegative and symmetric, has support equal to a compact interval, decreases to zero as a polynomial at the ends of its support, and has a bounded derivative; and $h \sim D_6\, n^{-C}$ as $n \to \infty$, where $C, D_6 > 0$.

Define $\hat\beta$ to be the minimizer of $S(\beta) = \sum_j (Y_j - \bar Y_j)^2$ (the first quantity in (2.7), corresponding to local-constant estimation) over functions $\beta = \sum_{1 \le k \le r} b_k\, \psi_k$, constrained by $\sum_{1 \le k \le r} b_k^2 = 1$, for which (3.5) and (3.6) hold and $\sup_{k \ge 1} |b_k| \le D_7$, with $D_7 > D_3$ (the latter as in (3.4)), and where

(3.8) r denotes the integer part of $D_8\, n^D$, for constants $D, D_8 > 0$.

This is the procedure for constructing $\hat g$ suggested in the argument leading to (2.8), in the local-constant case.

THEOREM 3.1. *If (3.1)–(3.8) hold, if B in (3.4) is sufficiently large, and if C and D in (3.7) and (3.8) are sufficiently small (all three constants depending only on $a_1, \ldots, a_4$), then there exists a constant $c > 0$ such that, as $n \to \infty$,*

$$n^{-1} \sum_{j=1}^{n} \{g(X_j) - \hat g(X_j)\}^2 = O_p(n^{-c})\,. \tag{3.9}$$

The proof is in Appendix A. It is possible to extend Theorem 1 to the recursive additive model case formulated in section 2.4, although the argument there is significantly longer. As explained above, for the case of Gaussian predictors X, the choices $a_1 = 1, a_2 = 1, a_3 = 1, a_4 = 1$ are possible and then by choosing the other constants judiciously, observing the various constraints, one finds that one may obtain the rate of convergence in (3.9) for c with $c < 1/4$. We do not pursue here the question of the optimality of this rate of convergence. An assumption that has been made throughout is that the

predictor trajectories are fully observed. This is an idealized situation. It is possible to weaken this assumption, assuming that the trajectories are sampled on a dense grid of points so that integrals such as those appearing in (2.12) can be closely approximated.

4. Algorithmic Implementation and Data Illustration.

4.1. *Description of the algorithm.*

Step 1. Estimating β. We assumed that h, r and the basis $\{\psi_1, \ldots, \psi_r\}$ (we used eigenbasis in our implementation) in (2.8) and (2.9) were given. We set $\hat{\beta} = \sum_{k=1}^{r} \hat{b}_k \psi_k$, and the coefficients $\hat{b}_1, \ldots, \hat{b}_r$ were estimated by minimizing (2.7). Those Y_j with $\sum_{i:i\neq j} K_{ij} < \lambda$ were dropped from the minimization (we chose $\lambda = 0.1$). Letting $\xi_{ik} = \int \psi_k x_i$ and writing $S(\beta)$ in (2.7) in terms of $b_1, \ldots, b_r$,

$$S(b_1, \ldots, b_r) = \frac{1}{n} \sum_{j=1}^{n} \left(Y_j - \sum_{i \neq j} w_{ij} Y_i \right)^2 \tag{4.1}$$

for the local-constant case, where

$$w_{ij}(b_1, \ldots, b_r, h) = \frac{K\left(h^{-1} \sum_{k=1}^{r} b_k(\xi_{ik} - \xi_{jk})\right)}{\sum_{l \neq j} K\left(h^{-1} \sum_{k=1}^{r} b_k(\xi_{lk} - \xi_{jk})\right)}$$

are the terms related to $b_1, \ldots, b_r$. For the local-linear case, $S(b_1, \ldots, b_r)$ is more complicated, with similar subsequent steps.

We note that $(b_1, \ldots, b_r, h)$ are not identifiable without constraints, since $w_{ij}(b_1, \ldots, b_r, h) = w_{ij}(cb_1, \ldots, cb_r, ch)$ for any constant c. Meanwhile, if K is symmetric, $w_{ij}(b_1, \ldots, b_r, h) = w_{ij}(-b_1, \ldots, -b_r, h)$. There are at least two ways to ensure algorithmic identifiability. In a first approach, given h, one may find $(b_1, \ldots, b_r)$ by minimizing (4.1), subject to the constraints $\sum_{k=1}^{r} b_k^2 = 1$ and $b_1 > 0$ (or $b_k > 0$ for some $b_k \neq 0$ if $b_1 = 0$). A second option is to find $(b_1, \ldots, b_r)$ that minimizes (4.1) near a suitable starting point $(c_1, \ldots, c_r)$, satisfying $\sum_{k=1}^{r} c_k^2 = 1$ and $c_1 > 0$, and then to rescale the solution to $(\frac{b_1}{\sqrt{\sum_k b_k^2}}, \ldots, \frac{b_r}{\sqrt{\sum_k b_k^2}}, \frac{h}{\sqrt{\sum_k b_k^2}})$. The second option is simpler since the unconstrained minimization is easier to achieve. However, if one wishes to specify h, the constraint $\sum_{k=1}^{r} b_k^2 = 1$ needs to be enforced in the minimization step. In the simulations, we found that both options led to virtually the same solution for a well-chosen bandwidth h.

The minimization step is a nonlinear least-squares problem, which can be implemented through the optimization package in MATLAB. It is important

to secure a good starting point for the minimization. We obtained a default starting point by searching along each dimension separately. Starting with the first dimension, we located a minimum along $S(b_1)$, as defined in (4.1), along a grid of values of b_1 in the interval $[0, 1]$. After obtaining the minimizer x_1, we continued to search along the second dimension using values $S(x_1, b_2)$, where b_2 varies on a grid within $[-1, 1]$. This approach was then iterated as necessary and provided the starting point.

Step 2. Selecting r and h. Here r is the number of eigenfunctions used in (2.8) and h is the kernel bandwidth. We employed 10-fold cross validation to evaluate each pair (h, r). Each of 10 subgroups of curves denoted by $V_1, \ldots, V_{10}$ was used as a validation set, one at a time, while the remaining data were used as the training set. For given (h, r), we found $\hat{\beta}$ as described in step 1 and computed $S(r, h) = \frac{1}{\sum_k \#V_k} \sum_{k=1}^{10} S_k$, where $S_k(r, h) = \sum_{j \in V_k} (Y_j - \hat{Y}_j)^2$ and $\hat{Y}_j = \hat{g}_1(\int \hat{\beta} X_j)$, using local-constant or local-linear method as described in the paragraph containing (2.9) and assuming only Y_i in the training set are known. We then found the minimizers of $S(r, h)$, which were the selected values for r and h.

Step 3. Backfitting step. By default, we fitted a single index functional regression model, which meant that predictions $\hat{g}(\int \hat{\beta}\, x_i)$ were obtained via (2.5) using the optimal (h, r) chosen in step 2 and the corresponding estimated β in step 1. For fitting a p-index functional regression model, the fits obtained in an initial single index step gave only $\hat{g}_1^0(\int \hat{\beta}_1^0\, x_i)$ in (2.13). We then replaced Y_i by $Y_i - \hat{g}_1^0(\int \hat{\beta}_1^0\, x_i)$ and repeated steps 1 and 2 to find $\hat{g}_2^0(\int \hat{\beta}_2^0\, x_i)$. This procedure was iterated until p indices was obtained. This only gives us the initial estimate of the p-index model. Then for the d-th iteration and the increasing sequence $k = 1, \ldots, p$, we uses $Y_i^d(k)$ defined in (2.14) to fit $\hat{g}_k^d(\int \hat{\beta}_k^d\, x)$. The iterations stops once $||\hat{\beta}_1^{d-1} - \hat{\beta}_1^d||_{L_2} < 0.01$ or 10 iterations are reached.

4.2. *Illustration for spectrometric data.* We applied the proposed model to spectrometric data that can be found at http://lib.stat.cmu.edu/datasets/tecator. We used only part of the data with data selection performed in the same way as in [3] and [15]. These data were obtained for 215 pieces of meat, for each of which one observes a spectrometric curve X_i, corresponding to an absorbance spectrum measured at 100 wavelengths. These spectrometric curves are depicted in Figure 1. The fat content of each sample was determined by an analytic method and recorded as a scalar response Y_i. One is interested in predicting the fat content of each sample directly from the

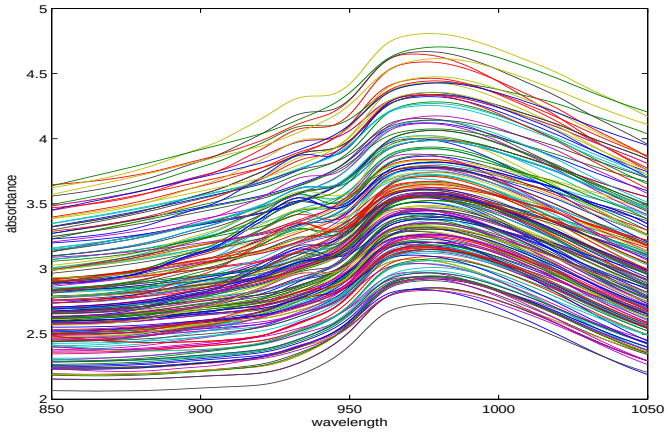


FIG 1. *Sample of* 204 *absorbance spectra for meat specimens.*

spectrometric curve.

In a preprocessing step, we removed 11 outliers. We also normalized each spectrometric curve by subtracting its area under the curve, $\int X_i(t)dt$, because we found that the first eigenfunction of the spectral curves is almost flat and its eigenvalue is much larger than the others, but the corresponding fitted coefficient $\hat{b}_1$ in (2.8) is close to 0. This normalization step reduced the leave-one-curve-out prediction error by more than 30%. The first four estimated eigenfunctions for the normalized curves are plotted in Figure 2.

To fit the functional single index model, we used 10-fold cross validation to choose the number r of included eigenfunctions in the representation (2.8) and the bandwidth for the Epanechnikov kernel, obtaining 4 and 0.0687 for these choices. Using the local linear method described in (2.5) and (2.7), we then estimated the regression parameter function β_1 and the link function g_1. These function estimates are shown in the upper panel of Figure 3. The average leave-one-curve-out squared prediction error for the proposed single index model is 3.51, while fitting a Generalized Functional Linear Model (GFLM) led to a prediction error of 4.99, showing substantial improvement for the proposed model.

We further applied the backfitting procedure described in Section 2.4 to check whether a multiple index functional model is more appropriate for these data than a single index model. The average leave-one-curve-out squared prediction errors were found to be 2.39 for the model with two indices and 2.42 for three indices. The estimated regression parameter func-

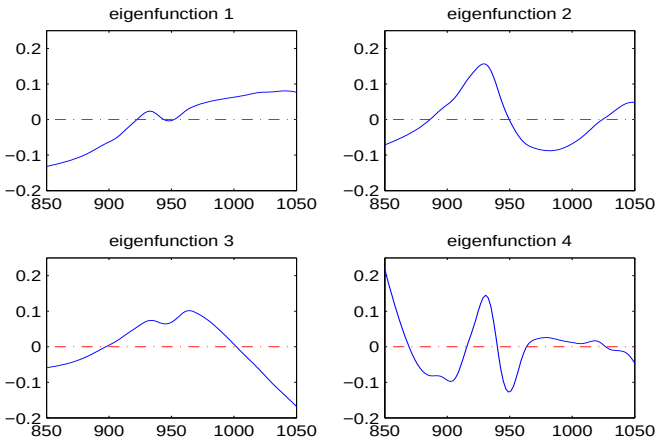


FIG 2. *The first four estimated eigenfunctions of the normalized absorbance spectra.*

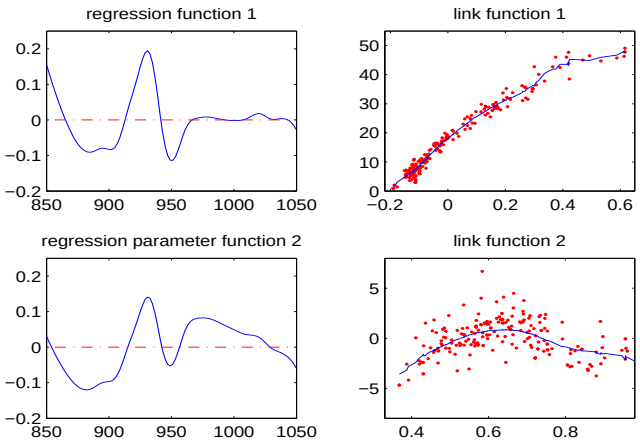


FIG 3. *The estimated regression parameter functions and link functions. Left two panels: the estimated regression parameter functions* $\hat{\beta}_1$ *and* $\hat{\beta}_2$ *for the first and second index, respectively; right two panels: the estimated link functions* $\hat{g}_1$ *and* $\hat{g}_2$ *for the first and second index, respectively.*

tions $\hat{\beta}_2$ and link function $\hat{g}_2$ are also displayed in Figure 3. The plot of $\hat{\beta}_1$ suggests that the small bump around wavelength 930 is an important indicator of the fat content level. We note that $\hat{\beta}_2$ has similar shape as $\hat{\beta}_1$ except for differences around wavelength 975, where it is positive. The model with two indices emerges as the best choice for prediction and improves more than 50% upon the GFLM and more than 30% upon the single index model in terms of prediction error.

5. Simulation Study.

5.1. *Simulations for single index models.* We studied the finite sample performance of five single index models (2.3). Samples of balanced functional data consisting of $N = 50/200/800$ predictor trajectories and a scalar response were generated and each predictor function was sampled through 50 equidistantly spaced measurements in $[0, 1]$. The predictor functions were generated as

$$X_i(t) = \mu(t) + \sum_{k=1}^{4} \xi_{ik}\phi_k(t), \quad i = 1, \ldots, N,$$

where $\mu(t) = t$, $\phi_1(t) = \frac{1}{\sqrt{2}}\sin(2\pi t)$, $\phi_2(t) = \frac{1}{\sqrt{2}}\cos(2\pi t)$, $\phi_3(t) = \frac{1}{\sqrt{2}}\sin(4\pi t)$, $\phi_4(t) = \frac{1}{\sqrt{2}}\cos(4\pi t)$, and ξ_{ik} are i.i.d. $N(0, \lambda_k)$ with $\lambda_1 = 1$, $\lambda_2 = \frac{1}{2}$, $\lambda_3 = \frac{1}{4}$, $\lambda_4 = \frac{1}{8}$. Responses Y_i were obtained as

Model i: $Y_i = \cos\left(\int_0^1 \beta\, X_i\right) + \epsilon_i$ (non-monotone link)

Model ii: $Y_i = \left(\int_0^1 \beta\, X_i\right)^2 + \epsilon_i$ (non-monotone link)

Model iii: $Y_i = \int_0^1 \beta\, X_i + \epsilon_i$ (functional linear model; trivially, aG monotone link)

Model iv: $Y_i \sim \text{Poisson}\left\{\exp\left(2 + \int_0^1 \beta\, X_i\right)\right\}$ (functional generalized Poisson model; a monotone link with heteroscedastic noise)

Model v: $Y_i \sim \text{Binomial}\left(1, \frac{1}{2}\cos\left(2\int_0^1 \beta\, X_i\right) + \frac{1}{2}\right)$ (functional generalized Binomial model; a non-monotone link with heteroscedastic noise)

where $\beta = \frac{1}{\sqrt{3}}\phi_1 + \frac{1}{\sqrt{3}}\phi_2 + \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$ in all models. In models (i)–(iii), errors ϵ_i were simulated as i.i.d Gaussian noise with mean 0 and $\text{var}(\epsilon) = R\,\text{var}\{g(\int \beta\, X)\}$. Here R is a measure of the signal-to-noise ratio, with values chosen as $R = 0.1$ and $R = 0.5$.

We compared the proposed model with the generalized functional linear regression model (GFLM) with unknown link and variance function [20],

TABLE 1
Simulation results for single index models (i)–(iii). 'FSIR' denotes the proposed functional single index regression and 'GFLM' denotes the generalized functional linear regression [20].

Model	N	FSIR				GFLM			
		R=0.1		R=0.5		R=0.1		R=0.5	
		RASE	RSE	RASE	RSE	RASE	RSE	RASE	RSE
i	50	.0464	.0204	.1096	.1225	.1299	.1488	.1546	.2335
	200	.0279	.0052	.0557	.0195	.0442	.0109	.0709	.0818
	800	.0156	.0024	.0315	.0041	.0288	.0025	.0402	.0049
ii	50	.1334	.0304	.3071	.2240	.1914	.1423	.3329	.3407
	200	.0731	.0065	.1549	.0223	.1058	.0150	.1838	.0840
	800	.0399	.0025	.0844	.0047	.0702	.0028	.0970	.0053
iii	50	.0970	.0341	.2562	.1705	.1024	.0546	.2378	.1819
	200	.0486	.0078	.1122	.0332	.0463	.0068	.1030	.0204
	800	.0226	.0030	.0526	.0083	.0237	.0026	.0558	.0071

TABLE 2
Simulation results for single index models (iv) and (v).

Model	N	FSIR		GFLM	
		RASE	RSE	RASE	RSE
iv	50	1.798	.0767	1.632	.0639
	200	1.207	.0214	1.064	.0137
	800	.8117	.0071	.6880	.0045
v	50	.2324	.4023	.2060	.4333
	200	.1222	.0850	.1400	.2866
	800	.0612	.0140	.0629	.0728

which is a single index model. In the simulations, we implemented the proposed model using the local-constant method defined in (2.4) (details can be found in section 4.1). Prediction outcomes were quantified by root average squared errors $\text{RASE} = \{\frac{1}{N}\sum_i\{\hat{Y}_i - g(\int \beta\, X_i)\}^2\}^{1/2}$, where $\hat{Y}_i$ is our estimate of $g(\int \beta\, X_i)$ defined in the paragraph containing (2.5), plugging in $\hat{\beta}$ and always leaving Y_i out of the sample when calculating $\hat{Y}_i$. We also quantified the error of the estimated regression parameter function by root squared error $\text{RSE}(\hat{\beta}) = \{\int(\hat{\beta} - \beta)^2\}^{1/2}$. Average values of RASE and RSE obtained from 100 Monte Carlo runs were then used to evaluate the procedures.

The results in tables 1 and 2 indicate that the proposed method works clearly better than GFLM for models (i), (ii) and (v), where the link function is non-monotone. For model (iii), the performance of the two methods was found to be similar. In this example, the effect of the monotone link function (here it is linear), which would have been expected to favor the GFLM, but this may be counteracted by the fact that the GFLM fits an unnecessarily

complex model in the case of homogeneous errors, as it also includes a non-parametric variance function estimation step. In model (iv), where the link is monotone and the noise is heteroscedastic, the GFLM not unexpectedly performs better, as it is able to target the heteroscedastic errors, improving efficiency of the estimates. Overall, it emerges that the proposed method is clearly preferable in situations where the link function is non-monotone.

5.2. *Simulations for multiple index models.* We simulated data for five multiple index models, using the same processes and settings as described in section 5.1. Three of the models (vi.-viii.) contain two indices and two models (ix.-x.) contain three indices, as follows:

Model vi: $Y_i = \cos\big(\int_0^1 \beta_1\, X_i\big) + 0.5 \sin\big(\int_0^1 \beta_2\, X_i\big) + \epsilon_i$, (two non-monotone link functions) where $\beta_1 = \frac{1}{\sqrt{3}}\phi_1 + \frac{1}{\sqrt{3}}\phi_2 + \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$, and $\beta_2 = \frac{1}{\sqrt{3}}\phi_1 - \frac{1}{\sqrt{3}}\phi_2 - \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$.

Model vii: $Y_i = \int_0^1 \beta_1\, X_i + \exp\big(0.5 \int_0^1 \beta_2\, X_i\big) + \epsilon_i$, (two monotone link functions) where $\beta_1 = \frac{1}{\sqrt{3}}\phi_1 + \frac{1}{\sqrt{3}}\phi_2 + \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$ and $\beta_2 = \frac{1}{\sqrt{3}}\phi_1 - \frac{1}{\sqrt{3}}\phi_2 - \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$.

Model viii: $Y_i = \int_0^1 \beta_1\, X_i + 0.5\big(\int_0^1 \beta_2\, X_i\big)^2 + \epsilon_i$, (one non-monotone link and one monotone link) where $\beta_1 = \frac{1}{\sqrt{3}}\phi_1 + \frac{1}{\sqrt{3}}\phi_2 + \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$ and $\beta_2 = \frac{1}{\sqrt{3}}\phi_1 - \frac{1}{\sqrt{3}}\phi_2 - \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$.

Model ix: $Y_i = \int_0^1 \beta_1\, X_i + \exp\big(0.5 \int_0^1 \beta_2\, X_i\big) + 0.5\big(\int_0^1 \beta_1\, X_i\big)^2 + \epsilon_i$, (three link functions) where $\beta_1 = \frac{1}{\sqrt{3}}\phi_1 + \frac{1}{\sqrt{3}}\phi_2 + \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$, $\beta_2 = \frac{1}{\sqrt{3}}\phi_1 - \frac{1}{\sqrt{3}}\phi_2 - \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$ and $\beta_3 = -\frac{1}{\sqrt{3}}\phi_1 + \frac{1}{\sqrt{3}}\phi_2 + \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$.

Model x: $Y_i = \int_0^1 \beta_1\, X_i + 0.5\big(\int_0^1 \beta_1\, X_i\big)^2 + 0.25\big(\int_0^1 \beta_1\, X_i\big)^3 + \epsilon_i$, (three link functions) where $\beta_1 = \frac{1}{\sqrt{3}}\phi_1 + \frac{1}{\sqrt{3}}\phi_2 + \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$, $\beta_2 = \frac{1}{\sqrt{3}}\phi_1 - \frac{1}{\sqrt{3}}\phi_2 - \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$ and $\beta_3 = -\frac{1}{\sqrt{3}}\phi_1 + \frac{1}{\sqrt{3}}\phi_2 + \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$.

We compared the results from the recursive fitting procedure and the backfitting procedure in terms of root average squared errors

$$\mathrm{RASE}_k = \Big\{ \frac{1}{N} \sum_i \Big\{ \sum_{j=1}^{k} \hat{g}_j\Big(\int \hat{\beta}_j\, X_i \Big) - \sum_{j=1}^{p} g_j\Big(\int \beta_j\, X_i \Big) \Big\}^2 \Big\}^{\frac{1}{2}},$$

for a p–index model when fitting the first k indices. It is of interest to include cases $k < p$ (not fitting a sufficient number of indices) and $k > p$ (overfitting the number of indices) and to determine whether the best results are obtained for the correct number of indices, which would suggest choosing the number of indices by fitting various numbers of indices and choosing

TABLE 3

Simulation results for multiple index model (vi)–(viii) with two underlying indices. $RASE_k^R$, $k = 1, 2, 3$, stands for root average errors using the recursive fitting procedure and k indexes and $RASE_k^I$ for the same errors obtained when using the iterative backfitting procedure. Shown are average results based on 100 *Monte Carlo runs.*

	N	R	RASE_1^R	RASE_2^R	RASE_3^R	RASE_1^I	RASE_2^I	RASE_3^I
vi	50	0.1	.2975	.1960	.1872	.2975	.1209	.1483
	200	0.1	.3003	.1357	.1059	.3003	.0645	.0854
	50	0.5	.3206	.2683	.2797	.3206	.2048	.2810
	200	0.5	.3051	.1778	.1754	.3051	.1235	.1427
vii	50	0.1	.2107	.2076	.2144	.2107	.1991	.2312
	200	0.1	.1311	.1131	.1372	.1311	.1070	.1247
	50	0.5	.4238	.3937	.4592	.4238	.3786	.4335
	00	0.5	.2507	.2329	.2719	.2507	.2267	.2848
viii	50	0.1	.4463	.3184	.3317	.4463	.2310	.2817
	200	0.1	.4061	.1496	.1594	.4061	.1016	.1279
	50	0.5	.4818	.4872	.5327	.4818	.4632	.4698
	200	0.5	.4304	.2502	.2910	.4304	.2107	.2412

TABLE 4

Recursive fitting results for models (ix) and (x) with three indices.

Model	N	R	RASE_1^R	RASE_2^R	RASE_3^R	RASE_4^R
ix	50	0.1	.5107	.3417	.3518	.3772
	200	0.1	.4791	.2196	.2132	.2183
	50	0.5	.5453	.4810	.5104	.5297
	200	0.5	.5161	.3324	.3329	.3504
x	50	0.1	.5107	.3417	.3518	.3772
	200	0.1	.4792	.2327	.2264	.2316
	50	0.5	.6631	.6461	.6111	.6418
	200	0.5	.5161	.3224	.3329	.3504

the number according to the model with the best fit. Here $\hat{g}_j$ and $\hat{\beta}_j$ are estimated using both recursive and backfitting procedures. Accordingly, if the underlying model, selected from models vi.-x., contains p indices, we calculated the values for RASE_k for $k = 1, \ldots, p + 1$.

As one can see from the results in tables 3, 4 and 5, the recursive fitting procedure often does not identify the right number of indexes and for nearly all fits produces larger RASE values, as compared to the iterative backfitting procedure. The iterative method thus emerges as the preferred method.

APPENDIX A: PROOF OF THEOREM 1

We describe the details of the proof by breaking it up into several steps.

TABLE 5
Iterative backfitting results for models (ix) and (x) with three indices.

Model	N	R	RASE_1^I	RASE_2^I	RASE_3^I	RASE_4^I
ix	50	0.1	.5107	.3272	.3009	.3395
	200	0.1	.4791	.2084	.1422	.1988
	50	0.5	.5453	.5372	.5518	.6095
	200	0.5	.5161	.3350	.3137	.3980
x	50	0.1	.5107	.3501	.3106	.3495
	200	0.1	.4792	.2063	.1808	.1860
	50	0.5	.6631	.6291	.5825	.6476
	200	0.5	.5161	.3372	.3129	.3248

Step 1. Upper bound on mean summed squared error. Define $\gamma_j = g(X_j) = g_1(\int_{\mathcal{I}} \beta^0 X_j)$ and

$$\text{(A.1)}\quad \bar{\gamma}_j = \Big(\sum_{i\,:\,i\neq j} \gamma_i K_{ij} \Big) \Big/ \sum_{i\,:\,i\neq j} K_{ij}\,, \quad \bar{\epsilon}_j = \Big(\sum_{i\,:\,i\neq j} \epsilon_i K_{ij} \Big) \Big/ \sum_{i\,:\,i\neq j} K_{ij}\,.$$

To express their dependence on β, through $K_{ij} = K_{ij}(\beta)$, we shall write $\bar{\gamma}_j$, $\bar{\epsilon}_j$ and $\bar{Y}_j$ as $\bar{\gamma}(\beta)$, $\bar{\epsilon}_j(\beta)$ and $\bar{Y}_j(\beta)$, respectively. In this notation, $S(\beta) = S_0 + S_1(\beta) + S_2(\beta) + 2\,S_3(\beta)$, where $S_0 = \sum_{1\le j\le n} \epsilon_j^2$ and does not depend on β,

$$\text{(A.2)}\qquad \S_1(\beta) = \sum_{j=1}^{n} \{\gamma_j - \bar{Y}_j(\beta)\}^2\,, \quad S_2(\beta) = \sum_{j=1}^{n} \{\gamma_j - \bar{\gamma}_j(\beta)\}\,\epsilon_j\,,$$
$$S_3(\beta) = \sum_{j=1}^{n} \bar{\epsilon}_j(\beta)\,\epsilon_j\,.$$

Furthermore, $S_1(\beta) = S_4(\beta) - 2\,S_5(\beta) + S_6(\beta)$, where

$$\text{(A.3)}\qquad S_4(\beta) = \sum_{j=1}^{n} \{\gamma_j - \bar{\gamma}_j(\beta)\}^2\,, \quad S_5(\beta) = \sum_{j=1}^{n} \{\gamma_j - \bar{\gamma}_j(\beta)\}\,\bar{\epsilon}_j(\beta)\,,$$
$$S_6(\beta) = \sum_{j=1}^{n} \bar{\epsilon}_j(\beta)^2\,,$$

with notations as in (A.1).

Let $\mathcal{B}_1 = \mathcal{B}_1(n)$ denote a class of functions β, and suppose we can prove that

$$\text{(A.4)}\qquad \sup_{\beta\in\mathcal{B}_1} |S_k(\beta)| = O_p(\lambda_n)\,, \quad \text{for} \quad k = 2, 3, 5, 6\,,$$

where λ_n denotes a sequence of positive constants. Then,

$$\begin{aligned} S_1(\hat\beta) &= S(\hat\beta) - \{S_0 + S_2(\hat\beta) + 2\,S_3(\hat\beta)\} \\ &\le S(\beta^0) - \{S_0 + S_2(\hat\beta) + 2\,S_3(\hat\beta)\} \\ &= S_1(\beta^0) + S_2(\beta^0) + 2\,S_3(\beta^0) - \{S_2(\hat\beta) + 2\,S_3(\hat\beta)\} \\ &= S_4(\beta^0) - 2\,S_5(\beta^0) + S_6(\beta^0) + S_2(\beta^0) \\ &\qquad\qquad + 2\,S_3(\beta^0) - \{S_2(\hat\beta) + 2\,S_3(\hat\beta)\} \\ &= S_4(\beta^0) + O_p(\lambda_n)\,, \end{aligned} \tag{A.5}$$

where the inequality follows from the fact that $\beta = \hat\beta$ minimizes $S(\beta)$, the final identity follows from (A.4) provided that β^0 and $\hat\beta$ are both in $\mathcal{B}_1(n)$, and all other identities in this string hold true generally.

Without loss of generality, the support of K is contained in the interval $[-1, 1]$ (see (3.7)). If in addition $|g_1(u) - g_1(v)| \le D_1\,|u - v|^{a_1}$ for all u and v (see (3.2)), then $|\gamma_j - \bar\gamma_j(\beta^0)| \le D_1\,h^{a_1}$ for all j, and therefore

$$S_4(\beta^0) \le n\,(D_1\,h^{a_1})^2\,. \tag{A.6}$$

Together, (A.5) and (A.6) imply that

$$\sum_{j=1}^{n} \{g(X_j) - \hat g(X_j)\}^2 = O_p(\lambda_n + n\,h^{2a_1})\,. \tag{A.7}$$

Step 2. Decomposition of each set $S_k(\beta)$ into two parts. Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ denote the set of explanatory variables, and for each $\beta \in \mathcal{B}_1$ let $\mathcal{J} = \mathcal{J}(\beta) \subseteq \mathcal{J}^0 \equiv \{1, \ldots, n\}$ denote a random set which satisfies

$$P\Big[\#\{\mathcal{J}^0 \setminus \mathcal{J}(\beta)\} > 2\,D_5\,n\,h^{a_4} \quad \text{for some} \quad \beta \in \mathcal{B}_1\Big] \to 0 \tag{A.8}$$

as $n \to \infty$, where a_4 is as in (3.6). (The set $\mathcal{J}$ will be $\mathcal{X}$-measurable.) Define $S_k^{\mathcal{J}}(\beta)$, for $2 \le k \le 6$, to be the version of $S_k(\beta)$ that arises if, in the definitions at (2.1) and (2.2), we replace summation over $1 \le j \le n$ by summation over $j \in \mathcal{J}$. Since g is bounded, and all moments of the error variables ϵ_i are finite (see (3.3)), then $\sup_{1\le i\le n} |Y_i| = O_p(n^\eta)$ with probability 1, for all $\eta > 0$. Therefore, in view of (A.8),

$$\max_{k=1,\ldots,6} \sup_{\beta\in\mathcal{B}_1} \big|S_k(\beta) - S_k^{\mathcal{J}}(\beta)\big| = O_p(n^{1+\eta}\,h^{a_4}) \quad \text{for all } \eta > 0\,. \tag{A.9}$$

Step 3. Determining $\mathcal{J}$ for which (A.8) holds. Define $T_j(\beta) = \sum_{i:\, i \neq j} K_{ij}$, recall that $f(\cdot \,|\, \beta)$ denotes the probability density of $\int_{\mathcal{I}} \beta X$, and put

$$\alpha_j(\beta) = h \int K(u)\, f\big(\textstyle\int_{\mathcal{I}} \beta X_j - hu \,|\, \beta\big)\, du\,.$$

Then,

$$\begin{aligned} E\{T_j(\beta) \,|\, X_j\} &= \alpha_j(\beta)\,, \\ \mathrm{var}\{T_j(\beta) \,|\, X_j\} &\le (n-1)\, h \int K(u)^2\, f\big(\textstyle\int_{\mathcal{I}} \beta X_j - hu \,|\, \beta\big)\, du \\ &\le n\, (\sup K)\, \alpha_j(\beta)\,. \end{aligned}$$

Moreover, $0 \le K_{ij} \le \sup K$. Therefore by Bernstein's inequality, if $0 < c_1 < 1$,

$$\begin{aligned} P\{T_j(\beta) &\le (1-c_1)\, n\, \alpha_j(\beta) \,|\, X_j\} \\ &= P\{n\, \alpha_j(\beta) - T_j(\beta) \ge c_1\, n\, \alpha_j(\beta) \,|\, X_j\} \\ &\le \exp\Big[- \frac{\{c_1\, n\, \alpha_j(\beta)\}^2/2}{(\sup K)\, \{n\, \alpha_j(\beta) + \frac{1}{3}\, c_1\, n\, \alpha_j(\beta)\}} \Big] \\ &= \exp\Big\{ - \frac{c_1^2\, n\, \alpha_j(\beta)}{2\, (\sup K)\, (1 + \frac{1}{3}\, c_1)} \Big\}\,. \end{aligned} \tag{A.10}$$

Hence, defining $\mathcal{J}(\beta)$ to be the set of all integers j such that $\alpha_j(\beta) \ge n^{-c_2}\, h$, where $0 < c_2 < 1$; and putting $C_2 = c_1^2/\{2\, (\sup K)\, (1 + \frac{1}{3}\, c_1)\}$; we obtain:

$$\sup_{j \in \mathcal{J}(\beta)} P\{T_j(\beta) \le (1-c_1)\, n\, \alpha_j(\beta) \,|\, X_j\} \le \exp\big(-C_2\, n^{1-c_2}\, h\big)\,.$$

Therefore, since $\mathcal{J}(\beta)$ contains no more than n elements, then

$$\begin{aligned} P\Big\{T_j(\beta) \le (1-c_1)\, n\, \alpha_j(\beta) \quad &\text{for some } j \in \mathcal{J}(\beta) \text{ and some } \beta \in \mathcal{B}_1\Big\} \\ &\le n\, (\#\mathcal{B}_1)\, \exp\big(-C_2\, n^{1-c_2}\, h\big)\,. \end{aligned}$$

Hence, provided

$$\#\mathcal{B}_1 = O\{n^{-C_3 - 1}\, \exp\big(C_2\, n^{1-c_2}\, h\big)\} \tag{A.11}$$

for some $C_3 > 0$, we have:

$$P\Big\{T_j(\beta) > (1-c_1)\, n\, \alpha_j(\beta) \quad \text{for all } j \in \mathcal{J}(\beta) \text{ and all } \beta \in \mathcal{B}_1\Big\} \to 1\,. \tag{A.12}$$

Note too that if a_3 and a_4 are as in (3.6), if K is supported on $[-1,1]$, and if

$$(\sup K)^{-1}\, n^{-c_2} \le D_4\, h^{a_3}\,, \tag{A.13}$$

then

$$\begin{aligned}\#\{\mathcal{J}^0 \setminus \mathcal{J}(\beta)\} &= \sum_{j=1}^n I\{\alpha_j(\beta) < n^{-c_2}\, h\} \\ &= \sum_{j=1}^n I\Big\{ \int_{\mathcal{I}} K(u)\, f\Big(\int_{\mathcal{I}} \beta X_j - hu \;\Big|\; \beta \Big)\, du < n^{-c_2} \Big\} \\ &\le \sum_{j=1}^n I_j\,,\end{aligned}$$

where

$$I_j = I_j(\beta) = I\Big\{ \sup_{|u| \le h} f\Big(\int_{\mathcal{I}} \beta X_j - u \;\Big|\; \beta \Big) < D_4\, h^{a_3} \Big\}\,.$$

The random variables $I_1, \ldots, I_n$ are independent and identically distributed, and, in view of (3.6), $\pi(\beta) \equiv P\{I_j(\beta) = 1\} \le D_5\, h^{a_4}$. Therefore, by Bernstein's inequality,

$$\begin{aligned}P\Big\{ \sum_{j=1}^n I_j(\beta) > 2\, D_5\, h^{a_4} \Big\} &\le P\Big[\sum_{j=1}^n \{I_j(\beta) - \pi(\beta)\} > D_5\, h^{a_4} \Big] \\ &\le \exp\Big[- \frac{(D_5\, h^{a_4})^2/2}{n\, \pi(\beta)\, \{1 - \pi(\beta)\} + \frac{1}{3}\, D_5\, h^{a_4}} \Big] \le \exp\big(- 3\, D_5\, n\, h^{a_4}/8 \big)\,.\end{aligned}$$

Hence, provided

$$\#\mathcal{B}_1 = o\big\{ \exp\big(3\, D_5\, n\, h^{a_4}/8\big) \big\}\,, \tag{A.14}$$

result (A.8) holds.

Step 4. Bound for $E_{\mathcal{X}}\{S_k^{\mathcal{J}}(\beta)^{2m}\}$ for $k = 2, 3, 5, 6$ and integers $m \ge 1$. Write $E_{\mathcal{X}}$ for expectation conditional on $\mathcal{X}$, let $Q = Q(\beta)$ denote the infimum of $\sum_{i\,:\,i \ne j} K_{ij}$ over all $j \in \mathcal{J}$, and put $\sigma^2 = E(\epsilon^2)$. Defining $L_{ij} = K_{ij}/(\sum_{i_1\,:\,i_1 \ne j} K_{i_1 j})$, taking $m \ge 1$ to be an integer, and using Rosenthal's inequality, we deduce that for a constant $A(m)$ depending only on m,

$$\begin{aligned}E_{\mathcal{X}}(\bar\epsilon_j^{2m}) &\le A(m) \Big\{ \sigma^{2m} \Big(\sum_{i\,:\,i \ne j} L_{ij}^2 \Big)^m + E(\epsilon^{2m}) \sum_{i\,:\,i \ne j} L_{ij}^{2m} \Big\} \\ &\le A(m) \Big\{ \big(\sigma^2\, Q^{-1}\, \sup K\big)^m + E\big(\epsilon^{2m}\big)\, Q^{-(2m-1)}\, (\sup K)^{2m-1} \Big\}\,.\end{aligned} \tag{A.15}$$

Therefore,

$$E_{\mathcal{X}}\{S_6^{\mathcal{J}}(\beta)^{2m}\} \le \Big\{ \sum_{j\in\mathcal{J}(\beta)} \big(E_{\mathcal{X}}|\bar\epsilon_j|^{2m}\big)^{1/(2m)} \Big\}^{2m} \tag{A.16}$$
$$\le A(m)\, n^{2m} \Big\{ \big(\sigma^2\, Q^{-1}\, \sup K\big)^m + E\big(\epsilon^{2m}\big)\, Q^{-(2m-1)}\, (\sup K)^{2m-1} \Big\}.$$

Moreover, if $|g| \le C_1$ then $S_4^{\mathcal{J}}(\beta) \le n\,(2C_1)^2$, and so, since $S_5^{\mathcal{J}}(\beta)^2 \le S_4^{\mathcal{J}}(\beta)\, S_6^{\mathcal{J}}(\beta)$, then

$$E_{\mathcal{X}}\{S_5^{\mathcal{J}}(\beta)^{2m}\} \le S_4^{\mathcal{J}}(\beta)^m \{E_{\mathcal{X}} S_6^{\mathcal{J}}(\beta)^{2m}\}^{1/2} \tag{A.17}$$
$$\le \{n\,(2C_1)^2\}^m \{E_{\mathcal{X}} S_6^{\mathcal{J}}(\beta)^{2m}\}^{1/2}.$$

More simply, if $|g| \le C_1$ then $S_4^{\mathcal{J}}(\beta) \le n\,(2C_1)^2$ and $\sum_j |\gamma_j - \bar\gamma_j(\beta)|^{2m} \le n\,(2C_1)^{2m}$, both uniformly in β. Therefore,

$$E_{\mathcal{X}}\{S_2^{\mathcal{J}}(\beta)^{2m}\} \le A(m) \Big[\{\sigma^2\, S_4^{\mathcal{J}}(\beta)\}^m + E\big(\epsilon^{2m}\big) \sum_{j\in\mathcal{J}(\beta)} |\gamma_j - \bar\gamma_j(\beta)|^{2m} \Big]$$
$$\le A(m)\, (2C_1)^{2m} \big\{ (n\,\sigma^2)^m + n\, E\big(\epsilon^{2m}\big) \big\}. \tag{A.18}$$

Recall that the support of K is contained in the interval $[-1, 1]$. Let N_1 denote the maximum, over values $j \in \mathcal{J}$, of the number of indices k such that $|\int_{\mathcal{I}} \beta\,(X_j - X_k)| \le h$. Then, the series $\sum_{i:\,i\ne j} K_{ij}$ has, for each j, at most N_1 nonzero terms. Array the values of $\int_{\mathcal{I}} \beta\, X_j$, for $j \in \mathcal{J}$, on the real line, and group them into consecutive blocks of indices j, each block (except for the last remnant block) containing just N_1 values. Index these blocks, from left to right along the line, from 1 to N_2, where N_2 equals $\lfloor (\#\mathcal{J})/N_1 \rfloor$ or $\lfloor (\#\mathcal{J})/N_1 \rfloor + 1$ and $\lfloor x \rfloor$ denotes the integer part of x. Choose one point $\int_{\mathcal{I}} \beta\, X_j$ from each even-indexed block, and remove those points from the respective blocks; and repeat this until all the points are removed from all the blocks. Record, for each pass through the N_2 blocks, the removed sequence $j_1, \ldots, j_\nu$ of indices. (On the first pass, ν will equal $\lfloor N_2/2 \rfloor$ or $\lfloor N_2/2 \rfloor + 1$, but on later passes, ν may be reduced in size.) Now repeat this for odd-indexed blocks. Denote by $j_{k1}, \ldots, j_{kM_k}$, for $1 \le k \le N$ say, the different sequences $j_1, \ldots, j_\nu$ that are obtained in this way. The set of all such sequences represents a (disjoint) partition of the integers in $\mathcal{J}$, and in particular, $M_1 + \ldots + M_N = n$. By construction, for each k the random variables $\epsilon_{j_{k1}}\, \bar\epsilon_{j_{k1}}, \ldots, \epsilon_{j_{kM_k}}\, \bar\epsilon_{j_{kM_k}}$ are independent, conditional on $\mathcal{X}$;

the random integers N and $M_1, \ldots, M_N$ are measurable in the sigma-field generated by $\mathcal{X}$; $N \le 2N_1$; and $\max_k M_k \le \lfloor (\#\mathcal{J})/(2N_1) \rfloor + 1$.

Since

$$\sum_{j \in \mathcal{J}(\beta)} \epsilon_j \, \bar{\epsilon}_j = \sum_{k=1}^{N} \left(\epsilon_{j_{k1}} \, \bar{\epsilon}_{j_{k1}} + \ldots + \epsilon_{j_{kM_k}} \, \bar{\epsilon}_{j_{kM_k}} \right)$$

then, for any integer $m \ge 1$ and an absolute constant $A(m) \ge 1$, depending only on m,

$$\begin{aligned}
E_{\mathcal{X}}\{S_3^{\mathcal{J}}(\beta)^{2m}\} &= E_{\mathcal{X}}\left\{ \left(\sum_{j \in \mathcal{J}(\beta)} \epsilon_j \, \bar{\epsilon}_j \right)^{2m} \right\} \\
&\le \left(\sum_{k=1}^{N} \left[E_{\mathcal{X}}\left\{ \left| \epsilon_{j_{k1}} \, \bar{\epsilon}_{j_{k1}} + \ldots + \epsilon_{j_{kM_k}} \, \bar{\epsilon}_{j_{kM_k}} \right|^{2m} \right\} \right]^{1/(2m)} \right)^{2m} \\
&\le A(m) \left(\sum_{k=1}^{N} \left[\left\{ \sum_{\ell=1}^{M_k} E_{\mathcal{X}}\left(\epsilon_{j_{k\ell}}^2 \, \bar{\epsilon}_{j_{k\ell}}^2 \right) \right\}^m + \sum_{\ell=1}^{M_k} E_{\mathcal{X}}\left(|\epsilon_{j_{k\ell}} \, \bar{\epsilon}_{j_{k\ell}}|^{2m} \right) \right]^{1/(2m)} \right)^{2m} \\
&\le A(m) \, N^{2m} \max_{1 \le k \le N} \left[\sigma^{2m} \left\{ M_k \max_{1 \le \ell \le M_k} E_{\mathcal{X}}\left(\bar{\epsilon}_{j_{k\ell}}^2 \right) \right\}^m \right. \\
&\qquad \left. + E\epsilon^{2m} \, M_k \max_{1 \le \ell \le M_k} E_{\mathcal{X}}\left(|\bar{\epsilon}_{j_{k\ell}}|^{2m} \right) \right].
\end{aligned}$$

Therefore, by (A.15),

$$\begin{aligned}
E_{\mathcal{X}}\{S_3^{\mathcal{J}}(\beta)^{2m}\} \le A(m)^2 \, N^{2m} \left\{ \left(\sigma^2 \, Q^{-1} \sup K \right)^m \max_{1 \le k \le N} M_k^m \right. \\
\left. + E(\epsilon^{2m}) \, Q^{-(2m-1)} \left(\sup K \right)^{2m-1} \max_{1 \le k \le N} M_k \right\}. \qquad \text{(A.19)}
\end{aligned}$$

The constant $A(m)$ in these bounds can be taken equal to $(A\, m/\log m)^m$, where $A > 1$ denotes an absolute constant [16, 19]. From this property, and results (A.16), (A.17), (A.18) and (A.19), and recalling that $N \le 2\, N_1$ and $M_k \le (n/2N_1) + 1$, we deduce that for a constant $C_4 > 1$,

$$\begin{aligned}
\sum_{k=2,3,5,6} E_{\mathcal{X}}\{S_k^{\mathcal{J}}(\beta)^{2m}\} &\le (m/\log m)^{m/2} \, (C_4 n)^{2m} \left\{ Q^{-m/2} + E(\epsilon^{2m}) \, Q^{(1/2)-m} \right\} \\
&\quad + (m/\log m)^{2m} \, C_4^m \left\{ (n\, N_1/Q)^m + E(\epsilon^{2m}) \, n \, (N_1/Q)^{2m-1} \right\}. \qquad \text{(A.20)}
\end{aligned}$$

The contributions to the left-hand side from $S_3^{\mathcal{J}}$ and $S_5^{\mathcal{J}}$ dominate, and so the right-hand side represents, in effect, $E_{\mathcal{X}}\{S_3^{\mathcal{J}}(\beta)^{2m}\} + E_{\mathcal{X}}\{S_5^{\mathcal{J}}(\beta)^{2m}\}$.

Step 5. Upper bounds to N_1 and Q^{-1}. Let $T_j^{[1]}(\beta)$ denote the version of $T_j(\beta)$ in the special case where $K \equiv 1$ on $[-1,1]$ and $K = 0$ elsewhere, and write $\alpha_j^{[1]}(\beta) = h \int_{|u|\le 1} f(\int_{\mathcal{I}} \beta X_j - hu \,|\, \beta)\, du$, representing the corresponding version of $\alpha_j(\beta)$. In this notation, $N_1 = N_1(\beta)$ equals the maximum, over j, of the values of $T_j^{[1]}(\beta)$ for $j \in \mathcal{J}(\beta)$. The argument leading to (A.10) now gives:

$$\begin{aligned} &P\Big\{T_j^{[1]}(\beta) > (1+c_1)\, n\, \alpha_j^{[1]}(\beta) \;\Big|\; X_j\Big\} \\ &\qquad = P\Big\{T_j^{[1]}(\beta) - n\, \alpha_j^{[1]}(\beta) \ge c_1\, n\, \alpha_j^{[1]}(\beta) \;\Big|\; X_j\Big\} \\ &\qquad \le \exp\Big[- \frac{\{c_1\, n\, \alpha_j^{[1]}(\beta)\}^2/2}{n\, \alpha_j^{[1]}(\beta) + \frac{1}{3}\, c_1\, n\, \alpha_j^{[1]}(\beta)} \Big] = \exp\Big\{ - \frac{c_1^2\, n\, \alpha_j^{[1]}(\beta)}{2\,(1 + \frac{1}{3}\, c_1)} \Big\}. \end{aligned}$$

The analogue of (A.12) in this setting is, assuming that (A.11) holds:

$$P\Big\{T_j^{[1]}(\beta) \le (1+c_1)\, n\, \alpha_j^{[1]}(\beta) \quad \text{for all } j \text{ and all } \beta \in \mathcal{B}_1\Big\} \to 1\,. \tag{A.21}$$

Since $\alpha_j^{[1]}(\beta) \le h \sup f(\cdot \,|\, \beta)$ then, using (3.5), we deduce from (A.21) that for a constant $C_5 > 0$,

$$P\{N_1(\beta) \le C_5\, nh \quad \text{for all } \beta \in \mathcal{B}_1\} \to 1\,. \tag{A.22}$$

Observe too that

$$\begin{aligned} Q(\beta)^{-1} = \Big\{ \inf_{j \in \mathcal{J}(\beta)} T_j(\beta) \Big\}^{-1} &\le \Big\{ (1-c_1) \inf_{j \in \mathcal{J}(\beta)} n\, \alpha_j(\beta) \Big\}^{-1} \\ &\le (1-c_1)^{-1}\, n^{c_2 - 1}\, h^{-1}\,, \end{aligned} \tag{A.23}$$

where the first identity is just the definition of Q; the second, in view of (A.12), holds uniformly in $\beta \in \mathcal{B}_1$, with probability converging to 1 as $n \to \infty$; and the third is a consequence of the definition of $\mathcal{J}(\beta)$ as the set of j for which $\alpha_j(\beta) \ge n^{-c_2}\, h$.

Step 6. Proof of uniform convergence to zero of $n^{-1}\, S_k^{\mathcal{J}}(\beta)$ for $k = 2, 3, 5, 6$. Incorporating the bounds at (A.21) and (A.22) into (A.20), and taking m to diverge polynomially fast in n, we deduce that, for constants

$C_6, C_7 > 1$, and with probability converging to 1 as $n \to \infty$,

$$\begin{aligned} s(m,n) &\equiv \sup_{\beta\in\mathcal{B}_1} \sum_{k=2,3,5,6} E_{\mathcal{X}}\{S_k^{\mathcal{J}}(\beta)^{2m}\} \\ &\le (m/\log m)^{m/2}\,(C_6 n)^{2m}\,\{(n^{c_2-1}/h)^{m/2} + E(\epsilon^{2m}) \\ &\qquad (n^{c_2-1}/h)^{m-(1/2)}\} + (m/\log m)^{2m}\,C_6^m\,\{n^{m(c_2+1)} \\ &\qquad + E(\epsilon^{2m})\,n^{(2m-1)c_2+1}\} \\ &\le (C_7\,n)^{2m}\,\{(mn^{c_2-1}/h)^{m/2} + (m^{2a_2}n^{c_2-1}/h)^m\} \\ &\qquad + (C_7\,n^2)^m\,\{(m^2\,n^{c_2-1})^m + (m^{2(a_2+1)}\,n^{2c_2-2})^m\}\,. \end{aligned} \tag{A.24}$$

where, to obtain the last inequality, we used the bound $E|\epsilon|^m \le (D_2\,m)^{a_2 m}$ in (3.3).

Choose c_2, and further positive constants C_8, C_9, c_3, c_4, c_5, such that

$$c_2 + 2c_3\,\max(1,a_2) + c_5 < 1 \quad \text{and} \quad 0 < c_4 < c_5\,. \tag{A.25}$$

Take m equal to the integer part of n^{c_3} and

$$C_8\,n^{-c_5} \le h \le C_9\,n^{-c_4}\,. \tag{A.26}$$

The constant $c_2 \in (0,1)$ was introduced immediately below (A.10), and, up to (A.25), was subject only to the conditions (A.13) and $0 < c_2 < 1$. For any given a_3 and c_2, no matter how small the latter, we can ensure that (A.13) holds merely by taking c_5 (and thence c_4), in (A.26), sufficiently small. Since the results below continue to hold no matter how small we choose c_5 (and c_4), then we can be sure that (A.13) is satisfied.

It follows from (A.24)–(A.26) that, with probability converging to 1 as $n \to \infty$,

$$\begin{aligned} s(m,n) \le (C_7\,n)^{2m}\,\Big\{&(n^{c_2+c_3+c_5-1})^m + (n^{c_2+2a_1c_3+c_5-1})^m \\ &+ (n^{c_2+2c_3-1})^m + (n^{2\{c_2+c_3(a_2+1)+c_2-1\}})^m\Big\} \le 4\,(C_7\,n^{c_6+1})^m\,, \end{aligned}$$

where $c_6 = \max\{c_2+c_3+c_5, c_2+2a_1c_3+c_5, c_2+2c_3, c_2+c_3(a_2+1)+c_2\} < 1$. Therefore, if $0 < c_7 < 1 - c_6$ and we put $c_8 = (1-c_6-c_7)/2 > 0$ then, by Markov's inequality,

$$\begin{aligned} P\Big\{n^{-1}\sup_{\beta\in\mathcal{B}_1}\sum_{k=2,3,5,6}|S_k^{\mathcal{J}}(\beta)| > n^{-c_8}\;\Big|\;\mathcal{X}\Big\} &\le 16^m\,(\#\mathcal{B}_1)\,s(m,n)\,n^{-2m(1-c_8)} \\ &\le 4\,(\#\mathcal{B}_1)\,(16\,C_7\,n^{-c_7})^m\,, \end{aligned} \tag{A.27}$$

where the inequalities hold with probability converging to 1 as $n \to \infty$. Hence, provided that

$$(\#\mathcal{B}_1)\,(16\,C_7\,n^{-c_7})^m \to 0\,, \tag{A.28}$$

the left-hand side of (A.27) converges in probability to zero as $n \to \infty$. It follows that the unconditional form of that probability also converges to zero, and hence that

$$P\Big\{n^{-1} \sup_{\beta\in\mathcal{B}_1} \max_{k=2,3,5,6} |S_k^{\mathcal{J}}(\beta)| > n^{-c_8}\Big\} \to 0\,. \tag{A.29}$$

Step 7. Completion. (A.9) and (A.29) we deduce that, for all $\eta > 0$,

$$n^{-1} \sup_{\beta\in\mathcal{B}_1} \max_{k=2,3,5,6} |S_k(\beta)| = O_p(n^\eta\,h^{a_4} + n^{-c_8}) = O_p(n^{\eta-c_4} + n^{-c_8})\,, \tag{A.30}$$

where we used (A.26) to derive the final identity. Therefore, (A.4) holds with $\lambda_n = n^{1-c_9}$ for any $c_9 \in (0, \max(c_4, c_8))$. Hence we may use this value of λ_n in (A.7), establishing that

$$n^{-1} \sum_{j=1}^n \{g(X_j) - \hat{g}(X_j)\}^2 = O_p(n^{-c})\,, \tag{A.31}$$

with $c = \min(c_9, 2a_1c_4)$, where the estimator $\hat{\beta}$ used to define $\hat{g}$ is obtained by minimizing $S(\beta) = \sum_j (Y_j - \bar{Y}_j)^2$ (the first quantity in (2.7)) over $\beta \in \mathcal{B}_1$. (We used (A.26) to simplify the term in h^{2a_1} in (A.7).)

During the proof above we imposed on the class $\mathcal{B}_1$ the assumption that $\beta^0 \in \mathcal{B}_1$ (see the discussion following (A.5)), and also three conditions — (A.11), (A.14) and (A.28) — on the size of the class. The latter three conditions hold if

$$\#\mathcal{B}_1 = O\{\exp(n^{c_{10}})\}\,, \tag{A.32}$$

provided $0 < c_{10} < \min(1 - c_2 - c_5, 1 - a_4c_5, c_3)$. (Recall from Step 6 that m equals the integer part of n^{c_3}.) By choosing c_5 smaller if necessary we can ensure that the upper bound here is strictly positive, and so $c_{10} > 0$.

Let $0 < c_{11} < c_{10}$ and $c_{12} > 0$, define $r = r(n)$ to be the integer part of $n^{c_{11}}$, and let D_3 be as in (3.4). Let r be as stipulated in (3.8), and write $\mathcal{B}_2$ for the class of functions $\beta = \sum_{1\le k\le r} b_k\,\psi_k$ such that each $|b_k| \le D_3$ for $1 \le k \le r$. Let $\mathcal{B}_3$ be the set of elements of $\mathcal{B}_2$ for which each b_k, for $1 \le k \le r$, is an integer multiple of $n^{-c_{12}}$. The number of elements of $\mathcal{B}_3$ is bounded above by a constant multiple of

$$(2D_3\,n^{c_{12}})^r \le \exp(\text{const.}\,n^{c_{11}}\,\log n) = o\{\exp(n^{c_{10}})\}\,. \tag{A.33}$$

Put $\mathcal{B}_1 = \mathcal{B}_3 \cup \{\beta^0\}$. Then (A.32) follows from (A.33).

The following three properties hold: (a) The lattice on which $\mathcal{B}_3$ is based can be made arbitrarily fine in a polynomial sense, by choosing c_{12} sufficiently large; (b) $E\|X\|^\eta < \infty$ for some $\eta > 0$ (see (3.3)); and (c) K has a bounded derivative (see (3.7)). Given $\beta = \sum_{1\le k\le r} b_k\,\psi_k \in \mathcal{B}_2$, let $\beta^{\text{approx}} = \sum_{1\le k\le r} b_k^{\text{approx}}\,\psi_k$ be the element of $\mathcal{B}_2$ defined by taking b_k^{approx} to be the lattice value nearest to b_k, for $1 \le k \le r$. Define $S(n)$ to equal the maximum, over $1 \le i, j \le n$, of $\|X_i - X_j\|$. Property (b) implies that $S_n = O_p(n^{c_{13}})$ for some $c_{13} > 0$. Using this property, and (a) and (c), it can be proved, by taking c_{12} sufficiently large, that for any given $c_{14} > 0$,

$$\begin{aligned}\sup_{\beta\in\mathcal{B}_2}\max_{1\le i,j\le n}\bigl|K_{ij}(\beta) - K_{ij}(\beta^{\text{approx}})\bigr| &= O_p\Bigl\{S(n)\,h^{-1}\sup_{\beta\in\mathcal{B}_2}\|\beta-\beta^{\text{approx}}\|\Bigr\}\\ &= O_p(n^{-c_{14}})\,. \qquad \text{(A.34)}\end{aligned}$$

From this result and the other properties of K in (3.7) it can be shown that (A.31) continues to hold if $\hat\beta$ in the definition of $\hat g$ is replaced by the minimizer of $S(\beta) = \sum_j (Y_j - \bar Y_j)^2$ over $\beta \in \mathcal{B}_4 = \mathcal{B}_2 \cup \{\beta^0\}$.

Call this result (R).

The desired result (3.9) follows from (R), except that the set $\mathcal{B}_4$ contains β^0 as an unusual, adjoined element. Hence there is, in theory, a possibility that $\hat\beta = \beta^0$; this could not happen if we were to restrict $\hat\beta$ to elements of $\mathcal{B}_2$, as required when defining the estimator $\hat g$ in (3.9). To appreciate that this does not cause any difficulty, let $\beta^1 = \sum_{1\le k\le r} b_k^0\,\psi_k$ denote the approximation to β^0 obtained by dropping all but the first r terms in the expansion $\beta^0 = \sum_{k\ge1} b_k^0\,\psi_k$. The argument leading to (A.34) can be used to prove that, for $c_{14} > 0$ chosen arbitrarily large, there exists a value of $B = B(c_{14})$, in the second part of (3.4), such that,

$$\max_{1\le i,j\le n}\bigl|K_{ij}(\beta^0) - K_{ij}(\beta^1)\bigr| = O_p\{S(n)\,h^{-1}\,\|\beta^0-\beta^1\|\} = O_p(n^{-c_{14}})\,.$$

Arguing as before, this leads to the conclusion that β^0 can be dropped from $\mathcal{B}_4$ without damaging result (R).

ACKNOWLEDGEMENTS

We wish to thank two reviewers for very helpful comments and suggestions.

REFERENCES

[1] Aguilera, A., Escabias, M. and Valderrama, M. (2008). Discussion of different logistic models with functional data. *Computational Statistics & Data Analysis* **53** 151–163.

[2] Ait-Saïdi, A., Ferraty, F., Kassa, R. and Vieu, P. (2008). Cross-validated estimations in the single-functional index model. *Statistics* **42** 475–494.
[3] Borggaard, C. and Thodberg, H. (1992). Optimal minimal neural interpretation of spectra. *Analytical Chemistry* **64** 545–551.
[4] Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *Annals of Statistics* **34**. 2159–2179.
[5] Cardot, H., Crambes, C., Kneip, A. and Sarda, P. (2007). Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics and Data Analysis* **51** 4832–4848.
[6] Chiou, J.-M. and Müller, H.-G. (2004). Quasi-likelihood regression with multiple indices and smooth link and variance functions. *Scandinavian Journal of Statistics* **31** 367–386.
[7] Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Annals of Statistics* **37** 35–72.
[8] Escabias, M. (2004). Principal component estimation of functional logistic regression: Discussion of two different approaches. *Journal of Nonparametric Statistics* **16** 365–384.
[9] Escabias, M., Aguilera, A. and Valderrama, M. (2007). Functional pls logit regression model. *Computational Statistics & Data Analysis* **51** 4891–4902.
[10] Ferraty, F., Goïa, A. and Vieu, P. (2007a). On the using of modal curves for radar waveforms classification. *Computational Statistics & Data Analysis* **51** 4878–4890.
[11] Ferraty, F., Goïa, A. and Vieu, P. (2007b). Nonparametric functional methods: new tools for chemometric analysis. *Statistical Methods for Biostatistics and Related Fields* Springer, Berlin 245–264
[12] Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis* **44** 161–173.
[13] Ferraty, F. and Vieu, P. (2004). Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination. *Journal of Nonparametric Statistics* **16** 111–125.
[14] Ferraty, F. and Vieu, P. (2006a). Functional nonparametric statistics in action. *The Art of Semiparametrics* Contrib. Statist. Physica-Verlag/Springer, Heidelberg 112-129
[15] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice.* Springer, Berlin.
[16] Hitczenko, P. (1990). Best constants in martingale version of Rosenthal's inequality. *Annals of Probability* **18** 1656–1668.
[17] James, G. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society* **64** 411–432.
[18] James, G. M. and Silverman, B. W. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association* **100** 565–576.
[19] Johnson, W., Schechtman, G. and Zinn, J. (1985). Best constants in moment inequalities for linear combinations of independent and exchangeable random variables. *Annals of Probability* **13** 234–253.
[20] Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics* **33** 774–805.
[21] Niang, S. (2002). Estimation de la densité dans un espace de dimension infinie: Application aux diffusions. *Comp. Rend. Acad. Sci.* **334** 213–216.

Department of Statistics,
University of California, Davis,
CA 95616, USA
E-mail: dchen@wald.ucdavis.edu
mueller@wald.ucdavis.edu

Department of Mathematics and Statistics,
University of Melbourne,
Parkville, VIC, 3010, Australia
E-mail: halpstat@ms.unimelb.edu.au