Biometrika (2013), **100**, 3, *pp.* 607–622
© 2013 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/ast004
Advance Access publication 26 March 2013

Continuously additive models for nonlinear functional regression

By HANS-GEORG MÜLLER

Department of Statistics, University of California, Davis, California 95616, U.S.A.
hgmueller@ucdavis.edu

YICHAO WU

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
wu@stat.ncsu.edu

and FANG YAO

Department of Statistics, University of Toronto, Toronto, Ontario M5S 3G3, Canada
fyao@utstat.toronto.edu

Summary

We introduce continuously additive models, which can be viewed as extensions of additive regression models with vector predictors to the case of infinite-dimensional predictors. This approach produces a class of flexible functional nonlinear regression models, where random predictor curves are coupled with scalar responses. In continuously additive modelling, integrals taken over a smooth surface along graphs of predictor functions relate the predictors to the responses in a nonlinear fashion. We use tensor product basis expansions to fit the smooth regression surface that characterizes the model. In a theoretical investigation, we show that the predictions obtained from fitting continuously additive models are consistent and asymptotically normal. We also consider extensions to generalized responses. The proposed class of models outperforms existing functional regression models in simulations and real-data examples.

Some key words: Berkeley growth study; Functional data analysis; Functional regression; Gene expression; Generalized response; Stochastic process; Tensor spline.

1. Introduction

Functional regression is a central methodology of functional data analysis and provides models and techniques for regression settings that include a random predictor function, a situation frequently encountered in the analysis of longitudinal studies and signal processing (Hall et al., 2001) or of continuous time-tracking data (Faraway, 1997), as well as in spectral analysis (Goutis, 1998). In such situations, functional regression models are used to assess the dependence of scalar outcomes on stochastic process predictors, where pairs of functional predictors and responses are observed for a sample of independent subjects. We consider here the case of functional regression with scalar responses, which may be continuous or of generalized type. For continuous responses, the functional linear model is a standard tool and has been thoroughly investigated;

see, e.g., Cai & Hall (2006), Cardot et al. (2007) and Crambes et al. (2009). However, the inherent linearity of this model is a limiting factor; often, the model is not flexible enough to adequately reflect more complex relations, motivating the development of more flexible functional regression models. To relate generalized responses to functional predictors, the generalized functional linear model (James, 2002; Escabias et al., 2004; Cardot & Sarda, 2005; Müller & Stadtmüller, 2005; Reiss & Ogden, 2010) is a common tool but is subject to similar limitations.

Previous extensions of functional linear regression include nonparametric functional models (Ferraty & Vieu, 2006) and functional additive models where centred predictor functions are projected on eigenfunctions of the predictor process and the model is then assumed to be additive in the resulting functional principal components (Müller & Yao, 2008). Here we pursue a different kind of additivity, which occurs in the time domain rather than in the spectral domain, and develop nonlinear functional regression models that are not dependent on a preliminary functional principal components analysis yet are structurally stable and not subject to the curse of dimensionality (Hall et al., 2009). Additive models (Friedman & Stuetzle, 1981; Stone, 1985) have been successfully used for many regression situations that involve continuous predictors and both continuous and generalized responses (Mammen & Park, 2005; Yu et al., 2008; Carroll et al., 2008).

The functional predictors we consider in our proposed time-additive approach to functional regression are assumed to be observed on their entire domain, usually an interval; therefore we are not dealing with a large-p situation. Instead, we seek a model that accommodates uncountably many predictors from the outset. Sparsity in the additive components, as considered, for example, by Ravikumar et al. (2009) for additive modelling in the large-p case, is not particularly meaningful in the context of a functional predictor. A more natural approach to overcoming the inherent curse of dimensionality, which we take here, is to replace sparsity by continuity when predictors are smooth infinite-dimensional random trajectories.

The extension of the standard additive model to the case of an infinite-dimensional rather than a vector predictor is based on two crucial features: the functional predictors are smooth over time, and the dimension of the time domain that defines the dimension of the additive model corresponds to the continuum. These observations motivate the replacement of sums of additive functions by integrals and the collection of additive functions characterizing traditional vector additive models by a smooth additive surface.

2. CONTINUOUSLY ADDITIVE MODELLING

The functional data that we consider here are associated with predictor functions X that correspond to the realization of a smooth and square-integrable stochastic process on a finite domain $\mathcal{T}$, with mean function $E\{X(t)\} = \mu_X(t)$, which are paired with responses Y. Thus the data consist of pairs (X_i, Y_i) that are independently and identically distributed as (X, Y), for $i = 1, \ldots, n$, where $X_i \in L^2(\mathcal{T})$ and $Y_i \in \mathbb{R}$. In this setting, we propose the continuously additive model

$$E(Y \mid X) = E(Y) + \int_{\mathcal{T}} g\{t, X(t)\}\,\mathrm{d}t, \tag{1}$$

for a bivariate smooth, i.e., twice differentiable, additive surface $g : \mathcal{T} \times \mathbb{R} \to \mathbb{R}$, which is required to satisfy $E[g\{t, X(t)\}] = 0$ for all $t \in \mathcal{T}$ for identifiability; see the Appendix for further details. A similar model has been considered by McLean et al. (2013).

Conceptually, continuous additivity emerges in the limit of a sequence of additive regression models as the time grid $\{t_1, \ldots, t_m\}$ in $\mathcal{T}$ becomes increasingly dense; the additive regression functions $f_j(\cdot)$ $(j = 1, \ldots, m)$ can be represented as $f_j(\cdot) = g(t_j, \cdot)$ where $E[g\{t_j, X(t_j)\}] = 0$.

Taking the limit $m \to \infty$ of the standardized additive models

$$E\{Y \mid X(t_1), \ldots, X(t_m)\} = E(Y) + \frac{1}{m}\sum_{j=1}^{m} g\{t_j, X(t_j)\}$$

and replacing the sum by an integral then yields the continuously additive model (1). Special cases of (1) include the following examples.

Example 1. Upon choosing $g\{t, X(t)\} = \beta(t)[X(t) - E\{X(t)\}]$, where β is a smooth regression parameter function, one obtains the familiar functional linear model

$$E(Y \mid X) = E(Y) + \int_{\mathcal{T}} \beta(t)[X(t) - E\{X(t)\}]\,\mathrm{d}t. \tag{2}$$

Example 2. Functional transformation models are of interest for non-Gaussian predictor processes. They are obtained by choosing $g\{t, X(t)\} = \beta(t)(\zeta\{X(t)\} - E[\zeta\{X(t)\}])$, where ζ is a smooth transformation of $X(t)$, which gives

$$E(Y \mid X) = E(Y) + \int_{\mathcal{T}} \beta(t)(\zeta\{X(t)\} - E[\zeta\{X(t)\}])\,\mathrm{d}t. \tag{3}$$

A special case is where $\zeta\{X(t)\} = X(t) + \eta(t)X^2(t)$ for some function η, which leads to

$$E(Y \mid X) = E(Y) + \int_{\mathcal{T}} \beta(t)[X(t) - E\{X(t)\}]\,\mathrm{d}t + \int_{\mathcal{T}} \eta(t)\beta(t)[X^2(t) - E\{X^2(t)\}]\,\mathrm{d}t;$$

this is a special instance of the functional quadratic model

$$E(Y \mid X) = \beta_0 + \int_{\mathcal{T}} \beta(t)X(t)\,\mathrm{d}t + \int_{\mathcal{T}}\int_{\mathcal{T}} \gamma(s,t)X(s)X(t)\,\mathrm{d}s\,\mathrm{d}t \tag{4}$$

(Yao & Müller, 2010), where γ is a smooth regression surface.

Example 3. The time-varying functional transformation model arises from the choice $g\{t, X(t)\} = \beta(t)[X(t)^{\alpha(t)} - E\{X(t)^{\alpha(t)}\}]$, where $\alpha(t) > 0$ is a smooth time-varying transformation function, which yields

$$E(Y \mid X) = E(Y) + \int_{\mathcal{T}} \beta(t)[X(t)^{\alpha(t)} - E\{X(t)^{\alpha(t)}\}]\,\mathrm{d}t.$$

Example 4. Extending (3), the functional regression might be determined by M different transformations $\zeta_j\{X(t)\}$ $(j = 1, \ldots, M)$ of predictors X, leading to the model

$$E(Y \mid X) = E(Y) + \sum_{j=1}^{M} \int_{\mathcal{T}} \beta_j(t)(\zeta_j\{X(t)\} - E[\zeta_j\{X(t)\}])\,\mathrm{d}t.$$

We also study an extension of continuously additive models to the case of generalized responses by including a link function h. With a variance function v, this extension is

$$E(Y \mid X) = h\left[\beta_0 + \int_{\mathcal{T}} g\{t, X(t)\}\,\mathrm{d}t\right], \quad \mathrm{var}(Y \mid X) = v\{E(Y \mid X)\} \tag{5}$$

for a constant β_0, under the constraint $E[g\{t, X(t)\}] = 0$ for all $t \in \mathcal{T}$. This extension is analogous to the following extension of the functional linear model to the case of generalized responses considered by James (2002) and Müller & Stadtmüller (2005):

$$E(Y \mid X) = h\left\{\beta_0 + \int_{\mathcal{T}} \beta(t)X(t)\,\mathrm{d}t\right\}, \quad \mathrm{var}(Y \mid X) = v\{E(Y \mid X)\}, \tag{6}$$

which is the commonly used generalized functional linear model. For binary responses, a natural choice for h is the expit function $h(x) = \exp(x)/\{1 + \exp(x)\}$, and a natural choice for v is the binomial variance function $v(x) = x(1 - x)$.

3. Prediction with continuously additive models

The continuously additive model (1) is characterized by the smooth additive surface g. For any sets of orthonormal basis functions $\{\phi_j\}$ on the domain $\mathcal{T}$ and $\{\psi_j\}$ on the range or truncated range of X, where such basis functions may be derived from B-splines, for example, one can find coefficients γ_{jk} such that the smooth additive surface g in (1) can be represented as

$$g(t, x) = \sum_{j,k=1}^{\infty} \gamma_{jk}\phi_j(t)\psi_k(x)$$

before standardization. Introducing truncation points p and q, the function g is then determined by the coefficients γ_{jk}, for $j = 1, \ldots, p$ and $k = 1, \ldots, q$, in the approximation model

$$E(Y \mid X) \approx E(Y) + \sum_{j,k=1}^{p,q} \gamma_{jk} \int_{\mathcal{T}} \phi_j(t)(\psi_k\{X(t)\} - E[\psi_k\{X(t)\}])\,\mathrm{d}t. \tag{7}$$

We assume throughout that the predictor trajectories X are fully observed or densely sampled. If predictor trajectories are observed with noise, or are less densely sampled, one may employ smoothing as a pre-processing step to obtain continuous trajectories. A simple approach that works well for the implementation of (7) is to approximate the smooth surface g by a step function which is constant on bins that cover the domain of g, choosing the basis functions ϕ_j and ψ_k as zero-degree splines. It is sometimes opportune to transform the trajectories $X(t)$ so as to narrow the range of their values.

Formally, we approximate g by a step function $g_{p,q}$ that is constant over bins:

$$g_{p,q}(t, x) = \sum_{j,k=1}^{p,q} \gamma_{jk} 1_{\{(t,x)\in B_{jk}\}}, \quad \gamma_{jk} = g(t_j, x_k),$$

where B_{jk} $(j = 1, \ldots p;\ k = 1, \ldots, q)$ is the bin defined by $[t_j - 1/(2p),\ t_j + 1/(2p)] \times [x_k - 1/(2q),\ x_k + 1/(2q)]$ for equidistant partitions of the time domain with midpoints t_j and of the range of X with midpoints x_j; in the following both domains are standardized to [0, 1].

Define

$$I_{jk} = \{t \in [0, 1] : \{t, X(t)\} \in B_{jk}\}, \quad Z_{jk} = Z_{jk}(X) = \int 1_{I_{jk}}(t)\,\mathrm{d}t. \tag{8}$$

With $\gamma = (\gamma_{11}, \ldots, \gamma_{p1}, \ldots, \gamma_{1q}, \ldots, \gamma_{pq})^{\mathrm{T}}$, a useful approximation under standardization is

$$\begin{aligned} E(Y \mid X) = E(Y) + \int_0^1 g\{t, X(t)\}\,\mathrm{d}t &\approx \theta_{p,q}(X, \gamma) \\ &= E(Y) + \sum_{j,k=1}^{p,q} \gamma_{jk}\{Z_{jk} - E(Z_{jk})\}. \end{aligned} \tag{9}$$

If g is Lipschitz continuous, this approximation is bounded as follows:

$$\begin{aligned} |E(Y \mid X) - \theta_{p,q}(X, \gamma)| &\leqslant \sup_{|t-t'|\leqslant 1/p,\, |x-x'|\leqslant 1/q} 2|g(t, x) - g(t', x')| \sum_{j,k=1}^{p,q} \int 1_{I_{jk}}(t)\,\mathrm{d}t \\ &= O\left(\frac{1}{p} + \frac{1}{q}\right), \end{aligned} \tag{10}$$

where we assume that $p = p(n) \to \infty$ and $q = q(n) \to \infty$ as $n \to \infty$; these bounds are uniform over all predictors X. From (9) and (10), for increasing sequences p and q, consider the sequence of approximating prediction models $\theta_{p,q}$:

$$E\{E(Y \mid X) - \theta_{p,q}(X, \gamma)\}^2 = E\left[E(Y \mid X) - E(Y) - \int_{\mathcal{T}} g\{t, X(t)\}\,\mathrm{d}t\right]^2 + O\{(pq)^{-1}\}. \tag{11}$$

For prediction with continuously additive models, it suffices to obtain the pq parameters γ_{jk} of the standardized approximating model $\theta_{p,q}(X, \gamma)$. For the case of generalized responses as in (5), the linear approximation (9) can be motivated analogously to (11), leading to the generalized linear model $h^{-1}\{E(Y \mid X)\} = \beta_0 + \sum_{j,k=1}^{p,q} \gamma_{jk}\{Z_{jk} - E(Z_{jk})\}$. We deal with the resulting generalized estimating equations (Wedderburn, 1974) by regularization with penalized weighted iterated least squares, penalizing against discrete second-order differences, approximating the smoothness penalty $\int\{\partial^2 g(t, x)/\partial t^2\}^2\,\mathrm{d}t + \int\{\partial^2 g(t, x)/\partial x^2\}^2\,\mathrm{d}x$ by

$$P_{\mathrm{S}}(\gamma) = \sum_{j,k=1}^{p,q} \{p^2(\gamma_{j-1k} - 2\gamma_{j,k} + \gamma_{j+1,k})^2 + q^2(\gamma_{jk-1} - 2\gamma_{j,k} + \gamma_{jk+1})^2\}.$$

This penalty works if g is at least twice continuously differentiable. If higher-order derivatives exist, corresponding higher-order difference quotients can be considered (Marx & Eilers, 1996).

Once a penalty P has been selected, defining $Z_X = \mathrm{vec}\{Z_{jk}(X)\}$ with Z_{jk} as above and elements ordered analogously to γ, and abbreviating Z_{X_i} to Z_i for the ith subject, predictions are obtained by determining the vector γ that minimizes

$$\sum_{i=1}^{n} (Y_i - Z_i\gamma)^2 + \lambda P(\gamma), \tag{12}$$

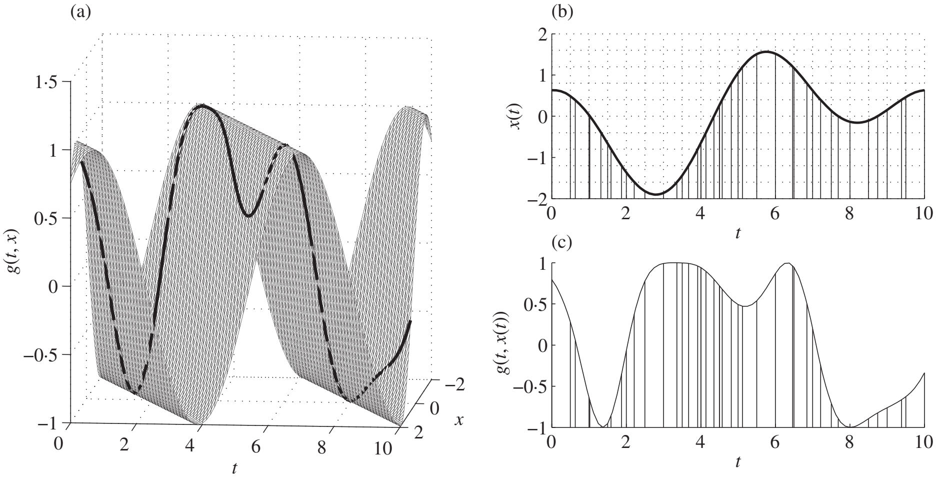

Fig. 1. Plots of the continuously additive model showing: (a) the smooth additive surface $g(t, x) = \cos(t - 5 - x)$; (b) a random function $X(t)$ from the sample; (c) $g\{t, X(t)\}$ as a function of t, for the random function X in (b).

followed by standardization. We determine the tuning parameter λ by K-fold crossvalidation. As the objective function (12) is quadratic for the penalties we consider, the computational aspects are straightforward.

For illustration, consider the smooth additive regression surface $g(t, x) = \cos(t - 5 - x)$ in Fig. 1(a), with domain of interest $t \in [0, 10]$ and $x \in [-2, 2]$. An example function $X(t)$ with overlaid bins B_{jk}, shown as a grid formed by dotted lines, is given in Fig. 1(b), where the value of Z_{jk} in (8) for each bin B_{jk} is defined by the distances between the solid vertical lines. The smooth additive surface g of Fig. 1(a), evaluated along the graph of the function X in Fig. 1(b), with $g\{t, X(t)\}$ viewed as a function of t, is displayed in Fig. 1(c), where the vertical lines are taken from Fig. 1(b). This serves to illustrate the approximation $\int_0^1 g\{t, X(t)\}\,\mathrm{d}t \approx \sum_{j,k=1}^{p,q} \gamma_{jk} Z_{jk}$ as in (9). Figure 1(a) also includes an example of the space curve $[t, X(t), g\{t, X(t)\}]$, parameterized in t, which is embedded in the smooth additive surface g and provides another visualization of the weighting that the graphs of predictor functions X are subjected to in the integration step leading to $E(Y \mid X) - E(Y) = \int g\{t, X(t)\}\,\mathrm{d}t$.

We are making the somewhat unrealistic assumption that entire predictor functions are observed. If this is not the case, or if one wishes to use derivatives of predictor functions, a common method is to presmooth discretely sampled and often noisy data. This approach has the advantage that it can be carried out for noisy measurements and somewhat irregularly spaced support points on which the functions are sampled. It is a common approach (Ramsay & Silverman, 2005) that leads to consistent representations of predictor trajectories under continuity and some additional regularity conditions, if designs are reasonably dense.

One can also extend the continuously additive model (1) to the case of multiple predictor functions by including one additive component of the type (1) for each predictor function, leading to a more complex approach that can be implemented analogously to the proposed methods. Other extensions of interest that can be relatively easily implemented and which may increase flexibility at the cost of more complexity include approximating the function g with higher-order spline functions and replacing the penalty $\lambda P(\gamma)$ in (12) by an anisotropic penalty employing two tuning parameters, such as $\sum_{j,k=1}\{\lambda_1(\gamma_{j-1k} - 2\gamma_{j,k} + \gamma_{j+1,k})^2 + \lambda_2(\gamma_{jk-1} - 2\gamma_{j,k} + \gamma_{jk+1})^2\}$.

4. Asymptotic properties

To study the asymptotic properties of predictors $\theta_{p,q}(X, \hat{\gamma})$ for $E(Y \mid X)$, where $\hat{\gamma}$ are the minimizers of the penalized estimating equations (12), we require the following assumptions in order to control the approximation error and to ensure that $\text{pr}\{\inf_{j,k} Z_{jk}(X) > 0\} > 0$.

Property 1. In both arguments t and x, $g : [0, 1]^2 \to \mathbb{R}$ is Lipschitz continuous.

Property 2. For all $t \in [0, 1]$, the random variable $X(t)$ has positive density on $[0, 1]$ and $X(\cdot)$ is continuous in t.

Quadratic penalties of the form $\gamma^{\mathrm{T}} P \gamma$ can be defined for semipositive-definite $pq \times pq$ penalty matrices P. For the specific penalty $P_{\mathrm{S}}(\gamma)$, one has the following additional characterizations. Using the $(l-2) \times l$ second-order difference operator matrix $D_l^2 a^{\mathrm{T}} = (a_1 - 2a_2 + a_3, \ldots, a_{l-2} - 2a_{l-1} + a_l)^{\mathrm{T}}$ for any $a \in \mathbb{R}^l$, the block diagonal matrix $\Delta_1 = \text{diag}(D_p^2, \ldots, D_p^2)$ with q such D_p^2 terms, and the matrix $P_1 = \Delta_1^{\mathrm{T}} \Delta_1$, the first term of $P_{\mathrm{S}}(\gamma)$ with index j can be written as $p^2 \gamma^{\mathrm{T}} P_1 \gamma$. Analogously, the second term with index k is $q^2 \gamma P_0^{\mathrm{T}} P_2 P_0 \gamma$; here P_0 is a permutation matrix with $P_0 \gamma = (\gamma_{11}, \ldots, \gamma_{1q}, \ldots, \gamma_{p1}, \ldots, \gamma_{pq})^{\mathrm{T}}$, and $P_2 = \Delta_2^{\mathrm{T}} \Delta_2$ where $\Delta_2 = \text{diag}(D_q^2, \ldots, D_q^2)$ with p such D_q^2 terms, so that $P_{\mathrm{S}} = p^2 P_1 + q^2 P_0^{\mathrm{T}} P_2 P_0$.

As the design matrix $Z = (Z_1, \ldots, Z_n)^{\mathrm{T}} = [\text{vec}\{Z_{jk}(X_1)\}, \ldots, \text{vec}\{Z_{jk}(X_n)\}]^{\mathrm{T}}$, with $Z_{jk}(X)$ as in (8), is not necessarily of full rank, A^{-1} in the following denotes the generalized inverse of a symmetric matrix A. If A admits a spectral decomposition $A = \sum_{\ell=1}^{s} \tau_\ell e_\ell e_\ell^{\mathrm{T}}$ with nonzero eigenvalues $\tau_1, \ldots, \tau_s$ and corresponding eigenvectors $e_1, \ldots, e_s$, where $s = \text{rank}(A)$, the generalized inverse is $A^{-1} = \sum_{\ell=1}^{s} \tau_\ell^{-1} e_\ell e_\ell^{\mathrm{T}}$. The spectral decomposition of $(Z^{\mathrm{T}} Z)^{-1/2} P (Z^{\mathrm{T}} Z)^{-1/2} = U D U^{\mathrm{T}}$ will be useful, where $D = \text{diag}(d_1, \ldots, d_{pq})$ is the diagonal matrix of nonincreasing eigenvalues, $d_1 \geqslant \cdots \geqslant d_r > d_{r+1} = \cdots = d_{pq} = 0$ with $r = \text{rank}(P)$, and U is the matrix of corresponding eigenvectors. For example, $r = pq - 2\min(p, q)$ for the second-order difference penalty P_{S}. With

$$\hat{\theta} = \{\theta_{p,q}(X_1, \hat{\gamma}), \ldots, \theta_{p,q}(X_n, \hat{\gamma})\}^{\mathrm{T}}, \quad \theta = \left[\int_0^1 g\{t, X_1(t)\}\, \mathrm{d}t, \ldots, \int_0^1 g\{t, X_n(t)\}\, \mathrm{d}t\right]^{\mathrm{T}},$$

the average mean square error, conditional on $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, is defined as

$$\text{AMSE}\,(\hat{\theta} \mid \mathcal{X}_n) = \frac{1}{n} E\{(\hat{\theta} - \theta)^{\mathrm{T}} (\hat{\theta} - \theta) \mid \mathcal{X}_n\}.$$

Theorem 1. *Assuming Properties 1 and 2, if $p \to \infty$, $q \to \infty$ and $pq/n \to 0$ as $n \to \infty$,*

$$\text{AMSE}\,(\hat{\theta} \mid \mathcal{X}_n) = O_{\mathrm{p}}\left\{\frac{1}{n} \sum_{\ell=1}^{pq} \frac{1}{(1 + \lambda d_\ell)^2} + \frac{\lambda^2}{n} \sum_{\ell=1}^{pq} \frac{d_\ell^2}{(1 + \lambda d_\ell)^2} + \frac{1}{pq}\right\}. \tag{13}$$

The first term on the right-hand side of (13) is due to variance, the second to shrinkage bias associated with the penalty, and the last to approximation bias. It is easy to see that the asymptotic variance and shrinkage bias trade off as λ varies, while a finer partition with larger p and q leads to decreased approximation bias.

To study the pointwise asymptotics at a future predictor trajectory x that is independent of $\{(X_i, Y_i) : i = 1, \ldots, n\}$, denote the estimate of $E(Y \mid X = x, \mathcal{X}_n)$ by $\hat{\theta}(x)$. With design matrix Z, let $R = Z^{\mathrm{T}} Z / n$ and $Z_x = \text{vec}\{Z_{jk}(x)\}$, and denote the smallest positive eigenvalue of R by $\rho_1 = \rho_1(n)$. In the smoothing literature the penalty λ/n is often used.

THEOREM 2. *If* $\lambda \to \infty$ *and* $\lambda/(n\rho_1) = o_p(1)$ *as* $n \to \infty$, *then*

$$E\{\hat{\theta}(x) \mid \mathcal{X}_n\} - \theta(x) = -\frac{\lambda}{n} Z_x^{\mathrm{T}} R^{-1} P\gamma\{1 + o_p(1)\} + O_p(1/p + 1/q), \tag{14}$$

$$\mathrm{var}\{\hat{\theta}(x) \mid \mathcal{X}_n\} = \frac{\sigma^2}{n} Z_x^{\mathrm{T}} R^{-1} Z_x\{1 + o_p(1)\}.$$

Suppose that, in addition, $\min(p^2, q^2)/n \to \infty$ *and* $\lambda^2/(n\rho_1^2) = O_p(1)$*; then, conditional on the design* $\mathcal{X}_n$,

$$\{\hat{\theta}(x) - \theta(x) - b_\lambda(x)\}/v(x)^{1/2} \longrightarrow N(0, 1) \tag{15}$$

in distribution, where $b_\lambda(x) = -n^{-1}\lambda Z_x^{\mathrm{T}} R^{-1} P\gamma$ *and* $v(x) = n^{-1}\sigma^2 Z_x^{\mathrm{T}} R^{-1} Z_x$.

The asymptotic bias in (14) includes a shrinkage bias, as reflected in the first term, and an approximation error, reflected in the second term. Shrinkage also induces a variance reduction, which is of higher order in comparison with $v(x)$; see (A2) in the Appendix. To attain asymptotic normality, the additional technical condition $\min(p^2, q^2)/n \to \infty$ renders the approximation bias negligible relative to the asymptotic standard error $v(x)^{1/2}$, while $\lambda^2/(n\rho_1^2) = O_p(1)$ ensures that the shrinkage bias $b_\lambda(x)$ does not dominate $v(x)^{1/2}$.

The presence of a nonnegligible bias term in the asymptotic normality result means that this result is of more theoretical than practical interest, as confidence intervals centred around the expected value do not coincide with the correct confidence intervals due to the presence of bias. While the asymptotic normality result (15) gives concise conditions for asymptotic convergence and a clear separation of the error into a variance part $v(x)$ and a bias part $b_\lambda(x)$, thus allowing further discernment of subcomponents such as shrinkage bias and approximation error, more extensive results on rates of convergence and theoretical justifications for inference remain open problems.

5. SIMULATION RESULTS

To assess the practical behaviour of the proposed continuously additive model (1), we used simulations to study the impact of grid size selection and of transformations, as well as the comparative performance of models where the data are generated according to the continuously additive model (1) or the functional linear model (2). In all scenarios, smooth predictor curves were generated by taking $X(t) = \sum_{k=1}^{4} \xi_k \phi_k(t)$ for $t \in [0, 10]$ with $\xi_1 = \cos(U_1)$, $\xi_2 = \sin(U_1)$, $\xi_3 = \cos(U_2)$ and $\xi_4 = \sin(U_2)$, where U_1 and U_2 are independent and identically distributed as Un$[0, 2\pi]$, and $\phi_1(t) = \sin(2\pi t/T)$, $\phi_2(t) = \cos(2\pi t/T)$, $\phi_3(t) = \sin(4\pi t/T)$ and $\phi_4(t) = \cos(4\pi t/T)$ with $T = 10$.

One such predictor curve is shown in Fig. 1(b). Separate training, tuning and test sets of sizes 200, 200 and 1000, respectively, were generated for each simulation run, where the tuning data were used to select the needed regularization parameters by minimizing the sum of squared prediction errors, separately for each method, and the predictor model was then fitted on the training set and evaluated on the test set. Performance was measured in terms of the average root mean squared prediction error RMSPE $= \{\sum_{i=1}^{1000}(Y_i - \hat{Y}_i)^2/1000\}^{1/2}$, using independent test sets of size 1000 and then averaging over 100 such test sets.

Simulation 1. We studied the effect of the number of grid points on the performance of the continuously additive model. Responses were generated according to $Y = \int_0^{10} \cos\{t - X(t) - 5\}\,\mathrm{d}t + \epsilon$, where $\epsilon \sim N(0, 1)$. The corresponding smooth additive surface is depicted in Fig. 1.

Table 1. *Root mean squared prediction errors obtained from investigating various grid sizes in Simulation* 1. *Standard deviations are given in parentheses*

$n_t = n_x$	5	10	20	40	80
RMSPE	1·138 (0·013)	1·039 (0·015)	1·030 (0·017)	1·029 (0·017)	1·030 (0·017)

RMSPE, root mean squared prediction error.

Table 2. *Root mean squared prediction errors obtained from investigating various signal-to-noise ratios as quantified by σ^2 in Simulation* 2, *an alternative functional nonlinear regression model in Simulation* 3, *and a functional linear model in Simulation* 4. *Standard deviations are given in parentheses*

	σ^2	FLM	FQM	FAM	CAM
	4	2·434 (0·018)	2·440 (0·022)	2·412 (0·041)	2·200 (0·056)
Simulation 2	1	1·723 (0·013)	1·728 (0·016)	1·645 (0·052)	1·156 (0·037)
	0·25	1·494 (0·011)	1·498 (0·014)	1·377 (0·057)	0·680 (0·035)
Simulation 3	1	9·828 (0·106)	5·810 (0·101)	9·568 (1·356)	1·119 (0·029)
Simulation 4	1	0·990 (0·007)	0·992 (0·008)	0·993 (0·010)	0·997 (0·011)

FLM, functional linear model; FQM, functional quadratic model; FAM, functional additive model; CAM, continuously additive model.

Denoting the number of equidistantly spaced grid points in directions t and x by n_t and n_x, respectively, we chose $n_t = n_x$. To demonstrate the effect of grid selection, we considered $n_t = n_x = 5, 10, 20, 40$ and 80. The means and standard deviations of RMSPE obtained over 50 simulations are reported in Table 1. The errors are seen to be larger for very small grid sizes, but once the grid size is above a minimal level, they remain roughly constant. The conclusion is that grid size does not have a strong impact, as long as very small grid sizes are avoided. Accordingly, we choose $n_t = n_x = 40$ for all other simulations and data analyses.

Simulation 2. With data generated in the same way as in Simulation 1 for model $Y = \int_0^{10} \cos[\pi\{t - X(t) - 5\}]\,\mathrm{d}t + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$, we compared the performance of the continuously additive model (1) with that of the functional linear model (2), the functional quadratic model (4) and the functional additive model, where one assumes an additive effect of the functional principal components $\{\xi_1, \xi_2, \ldots\}$ of predictor processes:

$$E(Y \mid X) = \beta_0 + \sum_{j=1}^{\infty} f_j(\xi_j). \tag{16}$$

In model (16) the f_j are smooth additive functions, standardized so that $E\{f_j(\xi_j)\} = 0$. In implementations, the sum is truncated at a finite number of terms (Müller & Yao, 2008). To explore the effect of signal-to-noise ratio, three different levels of σ^2 were selected. Tuning parameters were selected separately for each method by minimizing the sum of the squared prediction errors over the tuning set. The results in terms of RMSPE for 100 simulations can be found in Table 2, indicating that the continuously additive model has the smallest prediction errors. While the advantage of the continuously additive model over the other methods seems to persist across the table, it is more evident in situations with smaller signal-to-noise ratio.

Simulation 3. Results for the model $Y = \int_0^{10} t \exp\{X(t)\}\, \mathrm{d}t + \epsilon$ with $\epsilon \sim N(0, 1)$, proceeding as in Simulation 1, are given in Table 2. In this scenario, the continuously additive model has a prediction error that is considerably smaller than that of the other methods.

Simulation 4. The true underlying model was chosen as a functional linear model, so that one would expect the functional linear model to be the best performer. Responses were generated according to $Y = \int_0^{10} X(t) \cos\{2\pi(t - 5)\}\, \mathrm{d}t + \epsilon$ with $\epsilon \sim N(0, 1)$. The results in Table 2 indicate that, compared with the benchmark (the functional linear model), the loss associated with the continuously additive model and the other comparison methods is small.

To summarize, in many nonlinear functional settings, continuously additive modelling can lead to substantially better functional prediction than established functional regression models, while the loss in the case of an underlying functional linear model is quite small.

6. EXAMPLES

6·1. *Predicting pubertal growth spurts*

Human growth curves observed for a sample of children from various growth studies have been successfully analyzed with functional methodology (Kneip & Gasser, 1992). One is often interested in predicting future growth outcomes for a child when height measurements are available up to a certain age. We aim to predict the size of the pubertal growth spurt for boys, as measured by the size of the maximum in the growth velocity curve. As the growth spurt for boys in this study occurred after the age of 11·5 years, predictions were based on 17 height measurements made on a nonequidistant time grid before the age of 11·5 years for each of $n = 39$ boys in the Berkeley Growth Study (Tuddenham & Snyder, 1954).

Specifically, to obtain growth velocities for the ith boy from height measurements h_{ij} taken at ages s_j in a pre-processing step, we formed difference quotients $x_{ij} = (h_{i(j+1)} - h_{ij})/(s_{j+1} - s_j)$ and $t_{ij} = (s_j + s_{j+1})/2$ for $j = 1, \ldots, 30$, using all 31 measurements available per child from birth to age 18, and then applied local linear smoothing with a small bandwidth to each of the scatter-plots $\{(t_{ij}, x_{ij}),\ j = 1, \ldots, 30\}$ for $i = 1, \ldots, 39$. This yielded estimated growth velocity curves, which were then used to identify pubertal peak growth velocity. One subject whose data constituted an outlier was removed from the sample. For the prediction, we used continuous predictor trajectories obtained by smoothing the height measurements made before age 11·5, while excluding all subsequent measurements.

The estimated smooth additive surface g of model (1) that is uniquely obtained under the constraints $E\{g(t, X(t)\} = 0$ for all t is shown in Fig. 2; it reveals that prediction with the fitted model relies on a strong gradient after age 6, extending from a growth velocity of −4 cm/year to one of 6 cm/year, such that higher growth velocity in this time period is associated with predicting a more expressed pubertal growth spurt. This indicates that prediction in the fitted model relies on differentiating velocities in the age period 6–10 years and suggests that the intensity of a faint so-called mid-growth spurt (Gasser et al., 1984) affects the predicted size of the pubertal spurt. The predictive velocity gradient vanishes after age 10. As a cautionary note, these interpretations are intended merely to gain an understanding as to how the predictions are obtained within the fitted model and will depend on the type of constraint one selects for the identifiability of g.

To assess the predictive performance of various methods, we randomly split the data into training and test sets of sizes 30 and 8, respectively. We applied five-fold crossvalidation over each training set to tune the regularization parameters and then evaluated the root mean squared

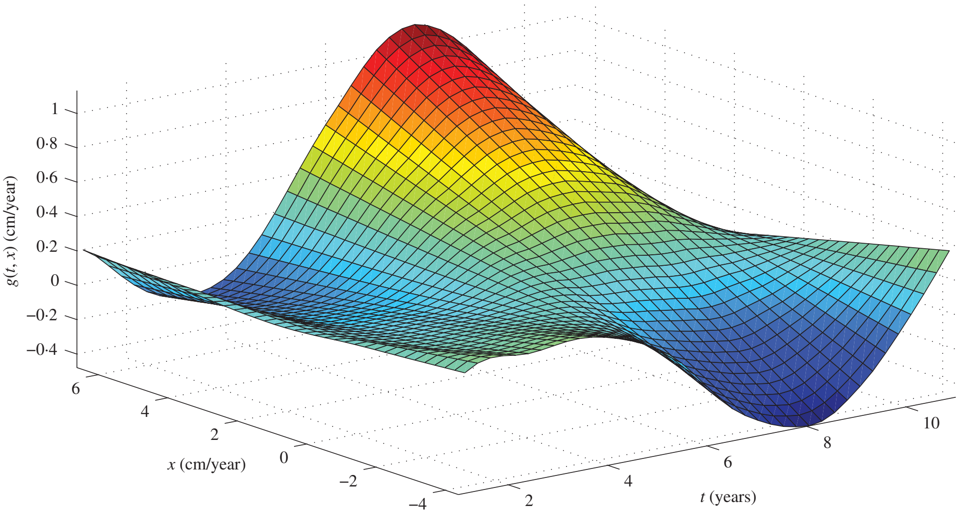

Fig. 2. Fitted smooth additive surface $g(t, x)$ for predicting pubertal growth spurts, obtained from one random partition of the data, for age t and growth velocity x.

Table 3. *Results of predicting pubertal growth spurts, comparing the root mean squared prediction errors and standard deviations (in parentheses)*

	FLM	FQM	FAM	CAM
RMSPE	0·549 (0·238)	0·602 (0·204)	0·606 (0·270)	0·502 (0·218)

RMSPE, root mean squared prediction error; FLM, functional linear model; FQM, functional quadratic model; FAM, functional additive model; CAM, continuously additive model.

prediction error over the test set; the results for ten random partitions are reported in Table 3. Among the methods compared, the continuously additive model was found to yield the best predictions of the intensity of the pubertal growth spurt.

6·2. *Classifying gene expression time courses*

We demonstrate use of the generalized version (5) of the continuously additive model for classification of gene expression time courses in brewer's yeast *Saccharomyces cerevisiae*; see Spellman et al. (1998) and Song et al. (2008) for details. Each gene expression time course features 18 gene expression measurements taken every 7 minutes, where the time origin corresponds to the beginning of the cell cycle. The task is to classify the genes according to whether they are related to the G1 phase regulation of the yeast cell cycle.

After removing an outlier, we used a subset of 91 genes with known classification and applied the continuously additive model (5) with a logistic link. The data were presmoothed and we used 40 uniform grid points over the domains of both t and x to obtain the fitted smooth additive surface $g(t, x)$, obtained under a constraint as before and shown in Fig. 3. At recording times near the left and right endpoints of the time domain, the gradient across increasing x is the highest, indicating that trajectory values near those endpoints have relatively large discriminatory power.

To assess classification performance, we considered, in addition to model (5) and analogously to model (1), versions of continuously additive models where the predictor processes X are transformed. The transformations include a simple timewise standardization transformation, where at each fixed time one subtracts the average trajectory value and divides by the standard deviation,

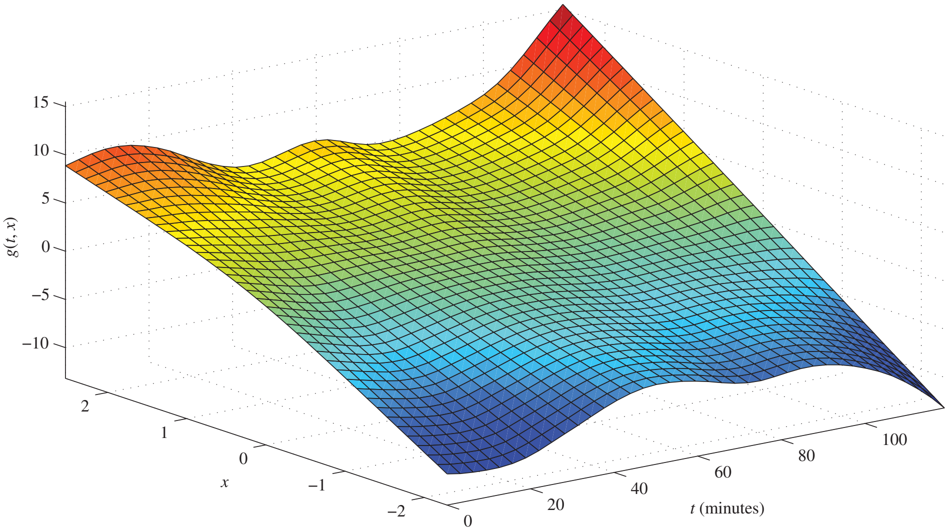

Fig. 3. Fitted smooth additive surface $g(t, x)$ for classifying yeast gene expression data, obtained from one random partition of the data, for time t and gene expression level x.

Table 4. *Results of classifying gene expression time courses for brewer's yeast, comparing average misclassification rates and standard deviations (in parentheses)*

GFLM	GCAM	GCAM-standardized	GCAM-range
0·156 (0·087)	0·097 (0·059)	0·097 (0·047)	0·144 (0·068)

GFLM, generalized functional linear model; GCAM, generalized continuously additive model; GCAM-standardized, generalized continuously additive model with a timewise standardization transformation applied; GCAM-range, generalized continuously additive model with a range transformation applied.

and a range transformation, where one standardizes for the range of observed values of $X(t)$ so that the range $\max X(t) - \min X(t)$ of the transformed predictors is invariant across all locations t. We also include a comparison with the generalized functional linear model (6).

For model comparisons, the 91 observations were randomly split into training and test sets of sizes 75 and 16, respectively. Tuning parameters were selected by five-fold crossvalidation in the training set, and models using these tuning parameters were fitted to the training data and then evaluated for the test data. We repeated the random split into training and test sets 20 times, and the average results for misclassification rates and standard deviations are reported in Table 4. We conclude that transformations do not necessarily improve upon the untransformed continuously additive model, and that the proposed model works better for this classification problem than does the generalized functional linear model.

Direct application of the functional quadratic model and functional additive model to the binary responses led to misclassification rates of 0·1344, standard deviation 0·0793, and 0·1531, standard deviation 0·0624, respectively. These results indicate that the proposed nonlinear functional regression model is competitive across a range of situations; this could be because it is more flexible than other existing functional regression models, while not being subject to the curse of dimensionality. The continuously additive model approach conveys, in a compact and interpretable way, the influence of the graph of the predictor trajectories on the outcome.

Acknowledgement

We wish to thank two reviewers and an associate editor for most helpful comments. In addition to several grants from the National Institutes of Health, the National Science Foundation and the National Science Research Council of Canada, the authors gratefully acknowledge support from the Statistical and Mathematical Sciences Institute at Triangle Park, North Carolina, where the bulk of this research was carried out within the framework of the Fall 2010 programme on Analysis of Object Data.

Appendix

Identifiability

Consider the unconstrained continuously additive model $E(Y \mid X) = \int_{\mathcal{T}} f\{t, X(t)\}\,\mathrm{d}t$. Here f is not identifiable. If the null space $\mathcal{N}(K)$ of the auto-covariance operator of predictor processes X satisfies $\mathcal{N}(K) = \{0\}$, then $\int_{\mathcal{T}} f\{t, X(t)\}\,\mathrm{d}t = 0$ with probability 1 implies that there is a one-dimensional function f^* on the domain $\mathcal{T}$ such that $f(t, x) \equiv f^*(t)$ and $\int_{\mathcal{T}} f^*(t)\,\mathrm{d}t = 0$.

As an example, consider the functional linear model, where g is linear in x or, more specifically, $g\{t, X(t)\} = \beta_0 + \beta(t)X(t)$. The intercept may be replaced with any function $\beta_0^*(t)$ such that $\int_{\mathcal{T}} \beta_0^*(t)\,\mathrm{d}t = \beta_0|\mathcal{T}|$, while the slope function $\beta(t)$ that is of primary interest is uniquely defined when $\mathcal{N}(K) = \{0\}$. More generally, one can express $g(t, x)$ with respect to x in a complete L^2 basis $\{1, x, x^2, \ldots, \}$, i.e., $g\{t, X(t)\} = \sum_{j=0}^{\infty} \beta_j(t)X^j(t)$. Then each $\beta_j(t)$ is uniquely defined. We can conclude that $g(t, x)$ is identifiable up to a function not depending on x.

The constraint $E[g\{t, X(t)\}] = 0$ for all $t \in \mathcal{T}$ thus ensures identifiability and also ties in with analogous constraints that are customarily made for the component functions in a conventional additive model with multivariate predictors, as well as for the functional linear model. The normalized model can be implemented by first obtaining an unconstrained f and then standardizing it via $g\{t, X(t)\} = f\{t, X(t)\} - E[f\{t, X(t)\}]$, where expectations are replaced by the corresponding sample means in implementations. The identifiability of the generalized version of the continuously additive model can be handled analogously.

Technical details

Proof of Theorem 1. Writing $Q = Z(Z^{\mathrm{T}}Z)^{-1/2}U$, it is easy to obtain the explicit solution $\hat{\gamma} = (Z^{\mathrm{T}}Z + \lambda P)^{-1}Z^{\mathrm{T}}Y$ where $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, for $\hat{\gamma}$ as in (12), and

$$\hat{\theta} = Z(Z^{\mathrm{T}}Z)^{-1/2}\{I + \lambda(Z^{\mathrm{T}}Z)^{-1/2}P(Z^{\mathrm{T}}Z)^{-1/2}\}^{-1}(Z^{\mathrm{T}}Z)^{-1/2}Z^{\mathrm{T}}Y = Q(I + \lambda D)^{-1}Q^{\mathrm{T}}Y.$$

With $Q^{\mathrm{T}}Q = U^{\mathrm{T}}U = I$ and the understanding that the following expectations are always conditional on $\mathcal{X}_n$ and therefore random, the covariance matrix of $\hat{\theta}$ is

$$\mathrm{var}(\hat{\theta}) = \sigma^2 Q(I + \lambda D)^{-1}Q^{\mathrm{T}}Q(I + \lambda D)^{-1}Q^{\mathrm{T}} = \sigma^2 Q(I + \lambda D)^{-2}Q^{\mathrm{T}},$$

which leads to

$$\frac{1}{n}E\{\|\hat{\theta} - E(\hat{\theta})\|^2\} = \frac{\sigma^2}{n}\mathrm{tr}\{(I + \lambda D)^{-2}Q^{\mathrm{T}}Q\} = \frac{\sigma^2}{n}\sum_{\ell=1}^{pq}\frac{1}{(1 + \lambda d_\ell)^2}.$$

To study the bias, denote the nonpenalized least-squares estimate by $\hat{\theta}_u = Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}Y = QQ^{\mathrm{T}}Y$ and observe that $E(\hat{\theta} - \theta) = E(\hat{\theta} - \hat{\theta}_u) + E(\hat{\theta}_u - \theta)$. Then

$$E(\hat{\theta} - \hat{\theta}_u) = Q\{(I + \lambda D)^{-1} - (I + \lambda D)^{-1}(I + \lambda D)\}Q^{\mathrm{T}}\theta = -\lambda Q(I + \lambda D)^{-1}DQ^{\mathrm{T}}\theta.$$

Since $Q^{\mathrm{T}}\theta = U^{\mathrm{T}}(Z^{\mathrm{T}}Z/n)^{-1/2}(Z^{\mathrm{T}}\theta/n^{1/2}) = O_{\mathrm{p}}(1)$ by the central limit theorem, one has

$$\frac{1}{n}\|E(\hat{\theta} - \hat{\theta}_u)\|^2 = O_{\mathrm{p}}\left[\frac{\lambda^2}{n}\mathrm{tr}\{D(I+\lambda D)^{-1}Q^{\mathrm{T}}Q(I+\lambda D)^{-1}D\}\right] = O_{\mathrm{p}}\left\{\frac{\lambda^2}{n}\sum_{\ell=1}^{pq}\frac{d_\ell^2}{(1+\lambda d_\ell)^2}\right\}.$$

For the approximation bias, writing $\theta_{p,q} = Z\gamma$ we have

$$E(\hat{\theta}_u - \theta) = Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}(Z\gamma + \theta - \theta_{p,q}) - \theta = (I - QQ^{\mathrm{T}})(\theta_{p,q} - \theta).$$

Using Property 1, the approximation error is $\|\theta_{p,q} - \theta\|_\infty = O(1/p + 1/q)$ from (11), and

$$\frac{1}{n}\|E(\hat{\theta}_u - \theta)\|^2 = O_{\mathrm{p}}\left\{\frac{1}{npq}\mathrm{tr}(I - QQ^{\mathrm{T}})\right\} = O_{\mathrm{p}}\left(\frac{n-pq}{npq}\right) = O_{\mathrm{p}}\left(\frac{1}{pq}\right). \qquad \square$$

Proof of Theorem 2. We first derive the asymptotic bias that consists of both approximation and shrinkage terms. The explicit solution is $\hat{\theta}(x) = Z_x^{\mathrm{T}}(Z^{\mathrm{T}}Z/n + \lambda P/n)^{-1}(Z^{\mathrm{T}}Y/n)$. For $R = Z^{\mathrm{T}}Z/n$, the maximum eigenvalue of $\lambda R^{-1}P/n$ is bounded by $c\lambda/(n\rho_1) = o_{\mathrm{p}}(1)$ for some constant c. Applying a Taylor expansion at $\lambda = 0$,

$$\begin{aligned}\hat{\theta}(x) &= Z_x^{\mathrm{T}}\left\{I - \frac{\lambda}{n}R^{-1}P + \left(\frac{\xi}{n}R^{-1}P\right)^2\right\}R^{-1}\frac{1}{n}Z^{\mathrm{T}}Y \\ &= \hat{\theta}_u(x) - \frac{\lambda}{n}R^{-1}PR^{-1}\frac{1}{n}Z^{\mathrm{T}}Y + r_n \qquad \text{(A1)}\end{aligned}$$

for some $\xi \in [0, \lambda]$, where $\hat{\theta}_u(x) = n^{-1}Z_x^{\mathrm{T}}R^{-1}Z^{\mathrm{T}}Y$ is the nonpenalized version and $r_n = Z_x^{\mathrm{T}}(\xi R^{-1}P)^2/n^3R^{-1}Z^{\mathrm{T}}Y$ is the remainder term. Since $\theta_{p,q} = Z\gamma$ and $\|\theta - \theta_{p,q}\|_\infty = O(1/p + 1/q)$, implying that $\theta_{p,q} - \theta = O_{\mathrm{p}}\{\theta_{p,q}(1/p + 1/q)\}$, we have

$$E\left(\frac{\lambda}{n}Z_x^{\mathrm{T}}R^{-1}PR^{-1}\frac{1}{n}Z^{\mathrm{T}}Y\right) = \frac{\lambda}{n}Z_x^{\mathrm{T}}R^{-1}P\gamma\{1 + o_{\mathrm{p}}(1)\}.$$

Analogously,

$$E(r_n) = \frac{\xi^2}{n^2}Z_x^{\mathrm{T}}(R^{-1}P)^2R^{-1}\frac{1}{n}Z^{\mathrm{T}}\theta = \frac{\xi^2}{n^2}Z_x^{\mathrm{T}}(R^{-1}P)^2\gamma\{1 + o_{\mathrm{p}}(1)\} = o_{\mathrm{p}}\left(\frac{\lambda}{n}Z_x^{\mathrm{T}}R^{-1}P\gamma\right).$$

For the approximation bias, writing $\theta_{p,q}(x) = n^{-1}Z_x^{\mathrm{T}}R^{-1}Z^{\mathrm{T}}\theta_{p,q} = Z_x^{\mathrm{T}}\gamma$ and noting that $|\theta_{p,q}(x) - \theta(x)| = O_{\mathrm{p}}(1/p + 1/q)$ from (11) and $\theta_{p,q} - \theta = O_{\mathrm{p}}\{\theta_{p,q}(1/p + 1/q)\}$ from above,

$$\begin{aligned}E\{\hat{\theta}_u(x) - \theta(x)\} &= \{\theta_{p,q}(x) - \theta(x)\} + Z_x^{\mathrm{T}}R^{-1}\frac{1}{n}Z^{\mathrm{T}}(\theta - \theta_{p,q}), \\ &= O_{\mathrm{p}}\left\{(1/p + 1/q)\left(1 + Z_x^{\mathrm{T}}R^{-1}\frac{1}{n}Z^{\mathrm{T}}Z\gamma\right)\right\} = O_{\mathrm{p}}(1/p + 1/q).\end{aligned}$$

For the asymptotic variance, with a Taylor expansion similar to (A1) we obtain

$$\begin{aligned}\mathrm{var}\{\hat{\theta}(x)\} &= \frac{\sigma^2}{n}Z_x^{\mathrm{T}}\left\{I - \frac{\lambda}{n}R^{-1}P + \left(\frac{\xi}{n}R^{-1}P\right)^2\right\}^2 R^{-1}Z_x \\ &= \frac{\sigma^2}{n}\left[Z_x^{\mathrm{T}}R^{-1}Z_x - \frac{2\lambda}{n}Z_x^{\mathrm{T}}R^{-1}PR^{-1}Z_x\{1 + o_{\mathrm{p}}(1)\}\right]. \qquad \text{(A2)}\end{aligned}$$

As the maximal eigenvalue of $\lambda R^{-1}/n$ is bounded from above by $\lambda/(n\rho_1) = o_{\mathrm{p}}(1)$, the second term in (A2) reflects a reduction of variance that corresponds to a higher-order term. Asymptotically, the leading term of the variance is $v(x) = \sigma^2 Z_x^{\mathrm{T}}R^{-1}Z_x/n$. Since, for an arbitrary X, $Z_X^{\mathrm{T}}Z_X = \sum_{\ell=1}^{pq} Z_{X,\ell}^2 \leqslant \sum_{\ell=1}^{pq} Z_{X,\ell} =$

1, the maximal eigenvalue of $R = n^{-1}\sum_{i=1}^{n} Z_{X_i} Z_{X_i}^{\mathrm{T}}$ is no greater than $\mathrm{tr}(R) \leqslant 1$, implying that $v(x)$ is bounded in probability by $\sigma^2 n^{-1} \leqslant v(x) \leqslant \sigma^2 (n\rho_1)^{-1}$.

To obtain asymptotic normality, conditional on the design $\mathcal{X}_n$, as $n \to \infty$ we require the approximation bias to be asymptotically negligible, i.e., $\min(p^2, q^2)v(x) \to \infty$. A sufficient condition is that $\min(p^2, q^2)/n \to \infty$. The shrinkage bias needs to satisfy $b_\lambda(x)/v(x)^{1/2} = O_{\mathrm{p}}(1)$, which is guaranteed by $\lambda^2/(n\rho_1^2) = O_{\mathrm{p}}(1)$. It remains to check the Lindeberg–Feller condition for the central limit theorem. As $v(x)$ and $\mathrm{var}\{\hat{\theta}(x)\}$ are asymptotically equivalent, it suffices to show that $[\hat{\theta}(x) - E\{\hat{\theta}(x)\}]/\mathrm{var}\{\hat{\theta}(x)\}^{1/2}$ converges to a standard normal distribution. Writing $\hat{\theta}(x) - E\{\hat{\theta}(x)\} = n^{-1} Z_x (R + n^{-1}\lambda P)^{-1} Z^{\mathrm{T}}(Y - \theta) = \sum_{i=1}^{n} a_i \epsilon_i$, where $a_i = n^{-1} Z_x (R + n^{-1}\lambda P)^{-1} Z_i$ and $\epsilon_i = Y_i - \theta(X_i)$, it will suffice to show that $\max_{1\leqslant i\leqslant n} a_i^2 = o_{\mathrm{p}}(\sum_{i=1}^{n} a_i^2) = o_{\mathrm{p}}[\mathrm{var}\{\hat{\theta}(x)\}]$.

Since the maximal eigenvalue of $Z_x Z_x^{\mathrm{T}}$ is no greater than $Z_x^{\mathrm{T}} Z_x \leqslant 1$ for any x, the fact that $\mathrm{tr}(AB) \leqslant \rho_A \mathrm{tr}(B)$ for any nonnegative-definite matrices A and B, where ρ_A is the maximal eigenvalue of A, implies that

$$a_i^2 = n^{-2} Z_x^{\mathrm{T}} \left(R + n^{-1}\lambda P\right)^{-1} Z_i Z_i^{\mathrm{T}} \left(R + n^{-1}\lambda P\right)^{-1} Z_x \leqslant n^{-2}\mathrm{tr}\left\{(I + n^{-1}\lambda R^{-1} P)^{-2} R^{-2} Z_i Z_i^{\mathrm{T}}\right\}.$$

By applying a Taylor expansion similar to the one above and observing that $\lambda/(n\rho_1) = o_{\mathrm{p}}(1)$, we see that the above quantity is bounded in probability by $(n\rho_1)^{-2}\mathrm{tr}(Z_i Z_i^{\mathrm{T}}) \leqslant (n\rho_1)^{-2}$. Then $\lambda^2/(n\rho_1^2) = O_{\mathrm{p}}(1)$ and $\lambda \to \infty$ imply that $\max_{1\leqslant i\leqslant n} a_i^2/v(x) \leqslant 1/(n\rho_1^2)$ converges to zero in probability, completing the proof. □

References

Cai, T. & Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34**, 2159–79.

Cardot, H., Crambes, C., Kneip, A. & Sarda, P. (2007). Smoothing splines estimators in functional linear regression with errors-in-variables. *Comp. Statist. Data Anal.* **51**, 4832–48.

Cardot, H. & Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *J. Mult. Anal.* **92**, 24–41.

Carroll, R., Aity, A., Mammen, E. & Yu, K. (2008). Nonparametric additive regression for repeatedly measured data. *Biometrika* **36**, 383–98.

Crambes, C., Kneip, A. & Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37**, 35–72.

Escabias, M., Aguilera, A. M. & Valderrama, M. J. (2004). Principal component estimation of functional logistic regression: Discussion of two different approaches. *J. Nonparam. Statist.* **16**, 365–84.

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics* **39**, 254–61.

Ferraty, F. & Vieu, P. (2006). *Nonparametric Functional Data Analysis.* New York: Springer.

Friedman, J. & Stuetzle, W. (1981). Projection pursuit regression. *J. Am. Statist. Assoc.* **76**, 817–23.

Gasser, T., Müller, H.-G., Köhler, W., Molinari, L. & Prader, A. (1984). Nonparametric regression analysis of growth curves. *Ann. Statist.* **12**, 210–29.

Goutis, C. (1998). Second-derivative functional regression with applications to near infra-red spectroscopy. *J. R. Statist. Soc.* B **60**, 103–14.

Hall, P., Müller, H.-G. & Yao, F. (2009). Estimation of functional derivatives. *Ann. Statist.* **37**, 3307–29.

Hall, P., Poskitt, D. S. & Presnell, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics* **43**, 1–9.

James, G. M. (2002). Generalized linear models with functional predictors. *J. R. Statist. Soc.* B **64**, 411–32.

Kneip, A. & Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist.* **20**, 1266–305.

Mammen, E. & Park, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *Ann. Statist.* **33**, 1260–94.

Marx, B. & Eilers, B. (1996). Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statist. Sci.* **11**, 89–121.

McLean, M. W., Hooker, G., Staicu, A. M., Scheipl, F. & Ruppert, D. (2013). Functional generalized additive models. *J. Comp. Graph. Statist.*, to appear.

Müller, H.-G. & Stadtmüller, U. (2005). Generalized functional linear models. *Ann. Statist.* **33**, 774–805.

Müller, H.-G. & Yao, F. (2008). Functional additive models. *J. Am. Statist. Assoc.* **103**, 1534–44.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*. New York: Springer, 2nd ed.

Ravikumar, P., Lafferty, J., Liu, H. & Wasserman, L. (2009). Sparse additive models. *J. R. Statist. Soc.* B **71**, 1009–30.

Reiss, P. & Ogden, R. (2010). Functional generalized linear models with images as predictors. *Biometrics* **66**, 61–9.

Song, J., Deng, W., Lee, H. & Kwon, D. (2008). Optimal classification for time-course gene expression data using functional data analysis. *Comp. Biol. Chem.* **32**, 426–32.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molec. Biol. Cell* **9**, 3273–97.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689–705.

Tuddenham, R. & Snyder, M. (1954). Physical growth of California boys and girls from birth to age 18. *Calif. Pub. Child Dev.* **1**, 183–364.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**, 439–47.

Yao, F. & Müller, H.-G. (2010). Functional quadratic regression. *Biometrika* **97**, 49–64.

Yu, K., Park, B. U. & Mammen, E. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.* **36**, 228–60.

[*Received April* 2012. *Revised January* 2013]