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, NC 27695, USA.

ywu11@ncsu.edu

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 motivated as extensions of additive regression models with vector predictors to the case of infinite-dimensional predictors. This approach provides a class of flexible functional nonlinear regression models, where random predictor curves are coupled with scalar responses. In continuously additive modeling, 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 approach outperforms existing functional regression models in simulations and data illustrations.

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 (FDA) 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), continuous time-tracking data (Faraway 1997) or 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 might 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; Crambes et al. 2009). However, the inherent linearity of this model is a limiting factor. This model is often not flexible enough to adequately reflect more complex functional regression 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 and is subject to similar limitations.

Previous extensions of functional linear regression include nonparametric functional models (Ferraty & Vieu 2006) or functional additive models where centered predictor functions are projected on eigenfunctions of the predictor process and then the model is assumed to be additive in the resulting functional principal components. (Müller & Yao 2008). Here we pursue a different kind of additivity that 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 component analysis yet are structurally stable and not weighed down by 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 the proposed time-additive approach to functional regression are assumed to be observed on their entire domain, usually an interval, and 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 in Ravikumar et al. (2009) for additive modeling in the large p case, is not particularly meaningful in the context of a functional predictor. A more natural approach to overcome the inherent curse of dimensionality is to replace sparsity by continuity when predictors are smooth infinite-dimensional random trajectories, as considered here.

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 across time and the dimension of the time domain that defines the dimension of the additive model corresponds to the continuum. These observations motivate to replace sums of additive functions by integrals and the collection of additive functions that characterizes traditional vector additive models by a smooth additive surface.

2. CONTINUOUSLY ADDITIVE MODELING

The functional data we consider here include 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 $E\{X(t)\} = \mu_X(t)$ and are paired with responses Y. The data then are pairs (X_i, Y_i) that are independently and identically distributed as (X, Y), $i = 1, \ldots, n$, where $X_i \in L^2(\mathcal{T})$ and $Y_i \in \mathfrak{R}$. For this setting, we propose the continuously additive model

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

for a bivariate smooth, i.e. twice differentiable, additive surface $g : \mathcal{T} \times \mathfrak{R} \to \mathfrak{R}$, which is required to satisfy $E\{g(t, X(t)\} = 0$ for all $t \in \mathcal{T}$ for identifiability; see Appendix for additional details.

Conceptually, continuous additivity emerges in the limit of a sequence of additive regression models for increasingly dense time grids $t_1, \ldots, t_m$ in $\mathcal{T}$, where additive regression functions $f_j(\cdot)$, $j = 1, \ldots, m$, can be represented as $f_j(\cdot) = g(t_j, \cdot)$ with $E\{g(t_j, X(t_j)\} = 0$ and taking the limit $m \to \infty$ in the standardized additive models

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

The continuously additive model (1) thus emerges by replacing sums by integrals in the limit. Special cases of the continuously additive model (1) include the following examples.

Example 1: The Functional Linear Model. Choosing $g\{t, X(t)\} = \beta(t)\{X(t) - EX(t)\}$, where β is a smooth regression parameter function, one obtains the familiar functional linear model

$$E(Y \mid X) = EY + \int_{\mathcal{T}} \beta(t)\{X(t) - EX(t)\}\, dt. \tag{2}$$

Example 2: Functional Transformation Models. Such models are of interest for non-Gaussian predictor processes and are obtained by choosing $g\{t, X(t)\} = \beta(t)[\zeta\{X(t)\} - E\zeta\{X(t)\}]$, where ζ is a smooth transformation of $X(t)$, leading to

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

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

$$E(Y \mid X) = EY + \int_{\mathcal{T}} \beta(t)\{X(t) - EX(t)\}dt + \int_{\mathcal{T}} \eta(t)\beta(t)\{X^2(t) - EX^2(t)\}\, dt$$

of the functional quadratic model (Yao & Müller 2010),

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

where γ is a smooth regression surface.

Example 3: The Time-Varying Functional Transformation Model. This model arises for the choice $g\{t, X(t)\} = \beta(t)\{X(t)^{\alpha(t)} - EX(t)^{\alpha(t)}\}$, where $\alpha(t) > 0$ is a smooth time-varying transformation function, yielding $E(Y \mid X) = EY + \int_{\mathcal{T}} \beta(t)\{X(t)^{\alpha(t)} - EX(t)^{\alpha(t)}\}\, dt$.

Example 4: The M-fold Functional Transformation Model. 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 $E(Y \mid X) = EY + \sum_{j=1}^{M} \int_{\mathcal{T}} \beta_j(t)[\zeta_j\{X(t)\} - E\zeta_j\{X(t)\}]\, dt$.

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

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

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 previously considered extension of the functional linear model to the case of generalized responses (James 2002; Müller & Stadtmüller 2005),

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

the commonly used generalized functional linear model. For binary responses, a natural choice for h is the expit function $h(x) = e^x/(1+e^x)$ and for v 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 orthonormal basis functions $\{\phi_j,\, j \geq 1\}$ on the domain $\mathcal{T}$ and $\{\psi_j,\, j \geq 1\}$ on the range or truncated range of X, where such basis functions for example may be derived from B-splines, one can find coefficients γ_{jk} such that the smooth additive surface g in (1) can be represented as

$$g(t, x) = \sum_{j,k=1}^{\infty} \gamma_{jk}\phi_j(t)\psi_k(x), \tag{7}$$

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

$$E(Y \mid X) \approx EY + \sum_{j,k=1}^{p,q} \gamma_{jk} \int_{\mathcal{T}} \phi_j(t)[\psi_k\{X(t)\} - E\psi_k\{X(t)\}]\,dt. \tag{8}$$

We assume throughout that 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 preprocessing step to obtain continuous trajectories. A simple approach that works well for the implementation of (8) is to approximate the smooth surface g by a step function, which is constant on bins that cover the domain of g, choosing basis functions ϕ_j and ψ_k as zero degree splines. It is sometimes opportune to transform the trajectories $X(t)$ to narrow the range of their values.

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

$$g_{pq}(t, x) = \sum_{j,k=1}^{p,q} \gamma_{jk} 1_{\{(t,x)\in B_{jk}\}}, \quad \gamma_{jk} = g(t_j, x_k), \tag{9}$$

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

$$I_{jk} = \{t \in [0,1] : \{t, X(t)\} \in B_{jk}\}, \quad Z_{jk} = Z_{jk}(X) = \int 1_{I_{jk}}(t)\, dt. \tag{10}$$

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

$$E(Y \mid X) = EY + \int_0^1 g\{t, X(t)\} dt \approx \theta_{p,q}(X, \gamma) = EY + \sum_{j,k=1}^{p,q} \gamma_{jk}(Z_{jk} - EZ_{jk}). \tag{11}$$

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

$$|E(Y \mid X) - \theta_{p,q}(X, \gamma)| \leq \sup_{|t-t'| \leq 1/p, |x-x'| \leq 1/q} 2|g(t,x) - g(t',x')| \sum_{j,k=1}^{p,q} \int 1_{I_{jk}}(t)\, dt = O(\frac{1}{p} + \frac{1}{q}), \tag{12}$$

where we assume that $p = p(n) \to \infty$, $q = q(n) \to \infty$ as $n \to \infty$; these bounds are uniform over all predictors X. From (11), (12), for increasing sequences p and q, consider the sequence of approximating prediction models $\theta_{p,q}$,

$$E\{E(Y \mid X) - \theta_{p,q}(X, \gamma)\}^2 = E[E(Y \mid X) - EY - \int_{\mathcal{T}} g\{t, X(t)\}\, dt]^2 + O(\frac{1}{pq}). \tag{13}$$

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 (11) may be motivated analogously to (13), leading to the generalized linear model $h^{-1}\{E(Y \mid X)\} = \beta_0 + \sum_{j,k=1}^{p,q} \gamma_{jk}(Z_{jk} - EZ_{jk})$. 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 $\int \{\partial^2 g(t,x)/\partial t^2\}^2\, dt + \int \{\partial^2 g(t,x)/\partial x^2\}^2\, dx$ by

$$P_S(\gamma) = \sum_{j,k=1}^{p,q} \{p^2(\gamma_{j-1k} - 2\gamma_{j,k} + \gamma_{j+1,k})^2 + q^2(\gamma_{jk-1} - 2\gamma_{j,k} + \gamma_{jk+1})^2\}. \tag{14}$$

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 $Z_X = \text{vec}\{Z_{jk}(X)\}$ with Z_{jk} as above and elements ordered analogously to γ, abbreviating Z_{X_i} by Z_i for the ith subject, predictions are obtained by

determining the vector γ that minimizes

$$\sum_{i=1}^{n}(Y_i - Z_i\gamma)^2 + \lambda P(\gamma), \tag{15}$$

followed by standardization. We determine the tuning parameter λ by K-fold cross-validation. As the objective function (15) 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 the left panel of Figure 1, 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 in the upper right panel, where the value of Z_{jk} in (10) for each bin B_{jk} is defined by the distances between the solid vertical lines. The smooth additive surface g of the left panel, evaluated along the graph of the function X in the upper right panel, viewing $g\{t, X(t)\}$ as a function of t, is displayed in the lower right panel, where the vertical lines are taken from the upper right panel. This serves to illustrate the approximation $\int_0^1 g\{t, X(t)\}\, dt \approx \sum_{j,k=1}^{p,q} \gamma_{jk} Z_{jk}$ as in (11). The left panel also includes a demonstration of the space curve $[t, X(t), g\{t, X(t)\}]$, parametrized 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 that leads to $E(Y \mid X) - EY = \int g\{t, X(t)\}\, dt$.

We make there the somewhat unrealistic assumption that entire predictor functions are observed. If this is not the case or 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 that 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 (15) by an anisotropic penalty, employing two tuning parameters such as $\sum_{j,k=1}\{\lambda_1(\gamma_{j-1k} - 2\gamma_{j,k} + \gamma_{j+1,k})^2 + \lambda_2(\gamma_{jk-1} - 2\gamma_{j,k} + \gamma_{jk+1})^2\}$.

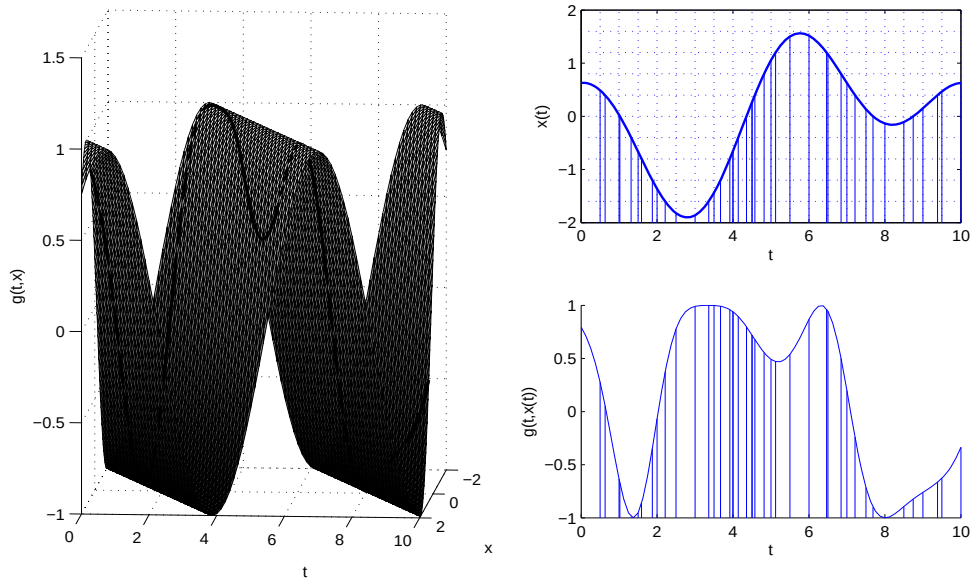

Figure 1: Illustrating the continuously additive model with the smooth additive surface $g(t,x) = \cos(t-5-x)$ (left), a random function $X(t)$ from the sample (upper right) and plotting $g\{t, X(t)\}$ as a function of t, for the random function X (lower right), also represented in the left panel.

4. ASYMPTOTIC PROPERTIES

To control the approximation error and to ensure that $\mathrm{pr}\{\inf_{j,k} Z_{jk}(X) > 0\} > 0$ for the study of 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 (15), we require

(A.1) $g : [0,1]^2 \to \Re$ is Lipschitz continuous in both arguments t and x.

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

While general quadratic penalties of the form $\gamma^\top P\gamma$ can be defined for semi-positive definite $pq \times pq$ penalty matrices P, one finds for the specific penalty $P_S(\gamma)$: Utilizing the $(l-2) \times l$ second order difference operator matrix matrix $D_l^2 a^\top = (a_1 - 2a_2 + a_3, \ldots, a_{l-2} - 2a_{l-1} + a_l)^\top$ for any $a \in \Re^l$, the block diagonal matrix $\Delta_1 = \mathrm{diag}(D_p^2, \ldots, D_p^2)$ with q such D_p^2 terms and the

matrix $P_1 = \Delta_1^\top \Delta_1$, the first term of $P_S(\gamma)$ with index j can be written as $p^2 \gamma^\top P_1 \gamma$. Analogously, the second term with index k is $q^2 \gamma P_0^\top P_2 P_0 \gamma$, where P_0 is a permutation matrix with $P_0\gamma = (\gamma_{11}, \ldots, \gamma_{1q}, \ldots, \gamma_{p1}, \ldots, \gamma_{pq})^\top$ and $P_2 = \Delta_2^\top \Delta_2$, $\Delta_2 = \mathrm{diag}(D_q^2, \ldots, D_q^2)$ with p such D_q^2 terms, so that $P_S = p^2 P_1 + q^2 P_0^\top P_2 P_0$.

As the design matrix $Z = (Z_1, \ldots, Z_n)^\top = [\mathrm{vec}\{Z_{jk}(X_1)\}, \ldots, \mathrm{vec}\{Z_{jk}(X_n)\}]^\top$, with $Z_{jk}(X)$ as in (10), 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 $A = \sum_{\ell=1}^s \tau_\ell e_\ell e_\ell^\top$ with nonzero eigenvalues $\tau_1, \ldots, \tau_s$ and corresponding eigenvectors $e_1, \ldots, e_s$, where $s = \mathrm{rank}(A)$, the generalized inverse is $A^{-1} = \sum_{\ell=1}^s \tau_\ell^{-1} e_\ell e_\ell^\top$. The spectral decomposition of $(Z^\top Z)^{-1/2} P (Z^\top Z)^{-1/2} = UDU^\top$ will be useful, where $D = \mathrm{diag}(d_1, \ldots, d_{pq})$ is the diagonal matrix of non-increasing eigenvalues, $d_1 \geq \ldots \geq d_r > d_{r+1} = \ldots = d_{pq} = 0$ with $r = \mathrm{rank}(P)$, 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

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

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

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

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

$$\mathrm{AMSE}\,(\hat{\theta} \mid \mathcal{X}_n) = O_p\left\{\frac{1}{n}\sum_{\ell=1}^{pq} \frac{1}{(1+\lambda d_\ell)^2} + \frac{\lambda^2}{n}\sum_{\ell=1}^{pq} \frac{d_\ell^2}{(1+\lambda d_\ell)^2} + \frac{1}{pq}\right\}. \tag{16}$$

The first term on the right hand side of (16) is due to variance, the second due to shrinkage bias associated with the penalty and the last due 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, 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 $R = Z^\top Z/n$, $Z_x = \mathrm{vec}\{Z_{jk}(x)\}$ and denote the smallest positive eigenvalue of R by $\rho_1 = \rho_1(n)$. Note that in the smoothing literature often the penalty λ/n is used.

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

$$E\{\hat{\theta}(x) \mid \mathcal{X}_n\} - \theta(x) = -\frac{\lambda}{n} Z_x^\top R^{-1} P\gamma\{1 + o_p(1)\} + O_p(1/p + 1/q) \quad (17)$$

$$\mathrm{var}\{\hat{\theta}(x) \mid \mathcal{X}_n\} = \frac{\sigma^2}{n} {Z_x}^\top R^{-1} Z_x\{1 + o_p(1)\}. \quad (18)$$

If 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* $\mathcal{X}_n$*,*

$$\{\hat{\theta}(x) - \theta(x) - b_\lambda(x)\}/\{v(x)\}^{1/2} \longrightarrow N(0,1) \quad \textit{in distribution}, \quad (19)$$

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

The asymptotic bias in (17) 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 (22) below. 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 non-negligible bias term in the asymptotic normality result means that this result is more of theoretical rather than practical interest, as confidence intervals centered around the expected value do not coincide with the correct confidence intervals due to the presence of bias. While the asymptotic normality result (19) provides concise conditions for the asymptotic convergence and a clear separation of the error into a variance part $v(x)$ and a bias part $b_\lambda(x)$, thus allowing to further discern subcomponents such as shrinkage bias and approximation error, further results on rates of convergence and theoretical justifications for inference remain open problems.

5. SIMULATION RESULTS

To assess the practical behavior of the proposed continuously additive model (1), we used simulations to study the impact of the grid size selection and of transformations and the comparative performance in models where the data are generated in conformance with the continuously additive model or with the functional linear model (2). In all scenarios, smooth predictor curves were generated according to $X(t) = \sum_{k=1}^{4} \xi_k \phi_k(t)$ for $t \in [0, 10]$ with $\xi_1 = \cos(U_1)$, $\xi_2 = \sin(U_1)$, $\xi_3 =$

$\cos(U_2)$, $\xi_4 = \sin(U_2)$, where U_1, U_2 are independent and identically distributed as Uniform$[0, 2\pi]$ and $\phi_1(t) = \sin(2\pi t/T)$, $\phi_2(t) = \cos(2\pi t/T)$, $\phi_3(t) = \sin(4\pi t/T)$, $\phi_4(t) = \cos(4\pi t/T)$ with $T = 10$.

One such predictor curve is shown in the upper right panel of Figure 1. Separate training, tuning, and test sets of sizes 200, 200, 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 SSPE, 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 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, 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 (CAM). Responses were generated according to $Y = \int_0^{10} \cos\{t - X(t) - 5\}dt + \epsilon$, where $\epsilon \sim N(0, 1)$. The corresponding smooth additive surface is depicted in Figure 1. 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 $n_t = n_x = 5, 10, 20, 40, 80$. The means and the corresponding standard deviations (in parentheses) 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 for CAM, as long as very small grid sizes are avoided. Accordingly, we choose $n_t = n_x = 40$ for all simulations and data analyses.

Table 1: Simulation results for root mean squared prediction error RMSPE for *Simulation 1*, investigating various grid selections. Standard deviations are in brackets.

$n_t = n_x$	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)

Simulation 2. Generating data in the same way as in Simulation 1 for model $Y = \int_0^{10} \cos[\pi\{t - X(t) - 5\}]\, dt + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$, we compared the performance of CAM (1) with that of the functional linear model (FLM) (2), the functional quadratic model (FQM) (4) and the functional additive model (FAM), where one assumes an additive effect of the functional principal

components $\{\xi_1, \xi_2, \ldots\}$ of predictor processes, i.e.

$$E(Y \mid X) = \beta_0 + \sum_{j=1}^{\infty} f_j(\xi_j). \tag{20}$$

In this model the f_j are smooth additive functions, standardized in such a way that $Ef_j(\xi_j) = 0$. 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 SSPE over the tuning set. The results in terms of RMSPE for 100 simulations can be found in Table 2, indicating that CAM has the smallest prediction errors. While the advantage of CAM over the other methods is seen to be persistent across the table, it is more expressed in situations with smaller signal to noise ratios.

Table 2: Simulation results for root mean squared prediction error RMSPE , 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 in brackets. Simulation comparisons are for the proposed continuously additive model CAMin comparison with the functional linear model FLM, functional quadratic model FQMand the functional additive model FAM.

Simulation	σ^2	FLM	FQM	FAM	CAM
Simulation 2	4	2.434 (0.018)	2.440 (0.022)	2.412 (0.041)	2.200 (0.056)
	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)

Simulation 3. Results for the model $Y = \int_0^{10} t \exp\{X(t)\}\, dt + \epsilon$ with $\epsilon \sim N(0,1)$, proceeding as in Simulation 1, are in Table 2. In this scenario CAM has a prediction error that is smaller by a large factor compared to the other methods.

Simulation 4. The true underlying model was chosen as a functional linear model, where one would expect FLM to be the best performer. Responses were generated according to $Y = \int_0^{10} X(t) \cos\{2\pi(t-$

$5)\}\, dt + \epsilon$ with $\epsilon \sim N(0,1)$. The results in Table 2 indicate that the loss of CAM and the other comparison methods compared to the benchmark FLM is small.

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

6. CONTINUOUSLY ADDITIVE MODELS IN ACTION

6.1 Predicting Pubertal Growth Spurts

Human growth curves observed for a sample of children in various growth studies have been successfully studied with functional methodology (Kneip & Gasser 1992). One is often interested to predict future growth outcomes for a child when height measurements are available up to a current 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 age 11.5 years, prediction was based on 17 height measurements made on a non-equidistant time grid before the age of 11.5 years for each of $n = 39$ boys in the Berkeley Growth Study (Tuddenham & Snyder 1954).

Specifically, for the ith boy, to obtain growth velocities from height measurements h_{ij} (in cm) at ages s_j in a preprocessing step, we formed difference quotients $x_{ij} = (h_{i(j+1)} - h_{ij})/(s_{j+1} - s_j),\ t_{ij} = (s_j + s_{j+1})/2$ for $j = 1, 2, \ldots, 30$, using all 31 measurements available per child between birth and 18 years, and then applied local linear smoothing with a small bandwidth to each of the scatterplots $\{(t_{ij}, x_{ij}),\ j = 1, 2, \ldots, 30\}$, $i = 1, \ldots, 39$. This yielded estimated growth velocity curves, which then were used to identify pubertal peak growth velocity. One subject with outlying data was removed from the sample. For the prediction, we used continuous predictor trajectories obtained by smoothing the height measurements made before age 11.5, 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 and is shown in Figure 2 reveals that prediction with the fitted model relies on a strong gradient after age 6, extending from growth velocity $x = -4$cm/yr to $x = 6$cm/yr, such that higher growth velocity in this time period is associated with predicting a

more expressed pubertal growth spurt. This indicates that the 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 merely intended 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.

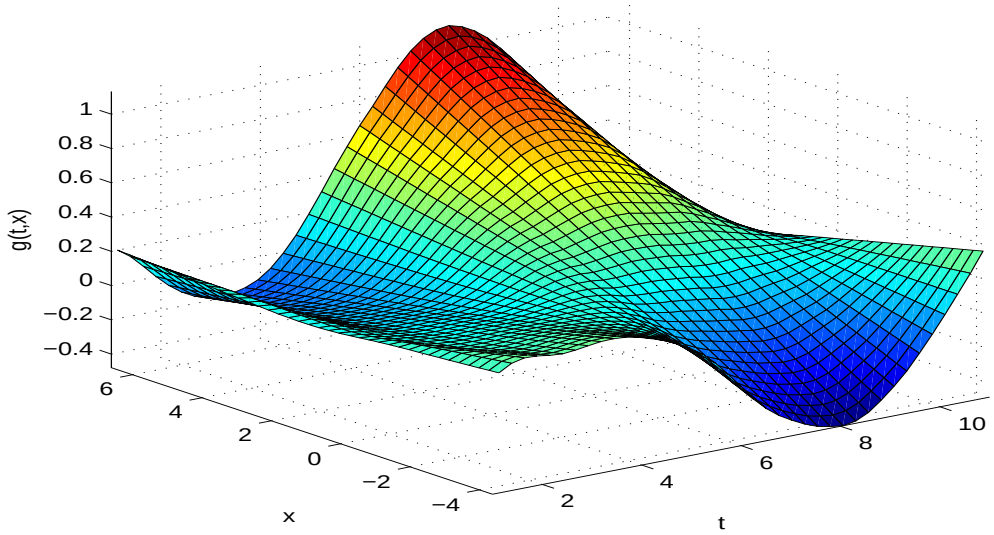

Figure 2: Fitted smooth additive surface $g(t, x)$ for predicting pubertal growth spurts as obtained for one random partition of the data. Age t is in years, growth velocity x in cm/year.

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 5-fold cross validation over the training set to tune the regularization parameters and then evaluated RMSPE over the test sets, with results for 10 random partitions reported in Table 3. Among the methods compared, CAM was found to yield the best predictions of the intensity of the pubertal growth spurt.

6.2 Classifying Gene Expression Time Courses

We demonstrate the generalized version of the continuously additive model (5) for the classification of yeast gene expression time courses for brewer's yeast (Saccharomyces cerevisiae); see Spellman et al. (1998); Song et al. (2008). Each gene expression time course features 18 gene ex-

pression measurements that have been taken every 7 minutes, where the origin of time 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.

Table 3: Results for predicting pubertal growth spurts, comparing root mean squared prediction errors RMSPE and standard deviations for functional linear model FLM(2), functional quadratic model FQM(4), functional additive model FAM(20) and continuously additive model CAM.

	FLM	FQM	FAM	CAM
RMSPE	0.549 (0.238)	0.602 (0.204)	0.606 (0.270)	0.502 (0.218)

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 both over the domains of t and of x to obtain the fitted smooth additive surface $g(t, x)$, as before obtained under a constraint and shown in Figure 3. At recording times near the left and right endpoints of the time domain the gradient across increasing x is highest, indicating that trajectory values near these endpoints have relatively large discriminatory power.

To assess classification performance, in addition to model (5) and analogously to model (1), we also considered versions of the continuously additive models where predictor processes X are transformed, including a simple timewise standardization transformation, where at each fixed time one subtracts the average trajectory value and divides by the standard deviation, and a range transformation, where one standardizes for the range of the observed values of $X(t)$, so that the range of the transformed predictors $\max X(t) - \min X(t)$ 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 sets of size 75 and test sets of size 16. Tuning parameters were selected by 5-fold cross-validation 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 these random splits 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 in comparison with the generalized functional linear model.

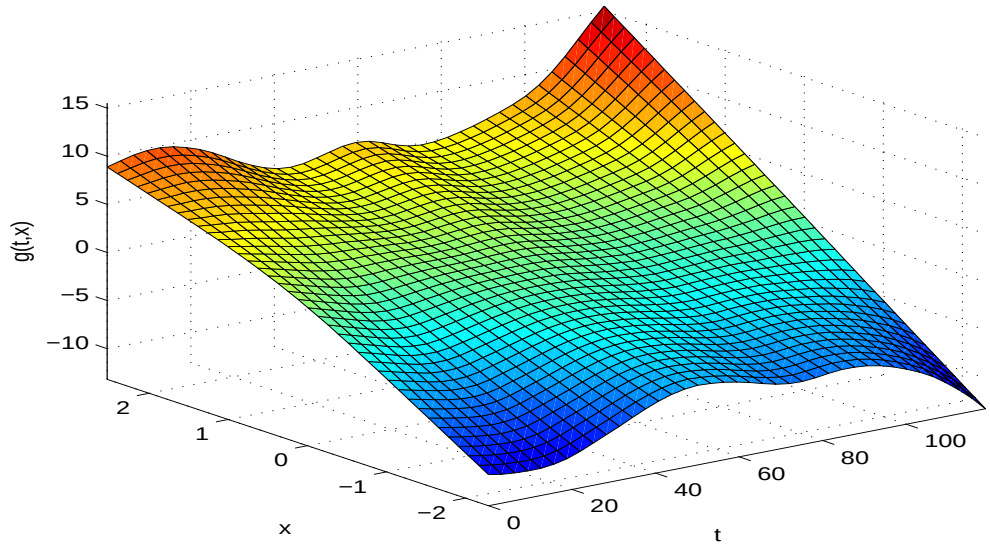

Figure 3: Fitted smooth additive surface $g(t, x)$ for classifying yeast gene expression data, as obtained for one random partition of the data, for gene expression level x and time t in minutes.

Table 4: Results for classifying gene time courses for brewer's yeast, comparing average misclassification rates AMRand standard deviations for generalized functional linear model GFLM(6), generalized continuously additive model GCAM (5) and GCAM combined with predictor transformation by standardization GCAM-standardized and with a range transformation GCAM-range.

	GFLM	GCAM	GCAM-standardized	GCAM-range
AMR	0.156(0.087)	0.097(0.059)	0.097(0.047)	0.144(0.068)

Direct application of FQM and FAM to the binary responses led to misclassification rates of 0.1344 (0.0793) for FQM and 0.1531 (0.0624) for FAM. These results indicate that the proposed nonlinear functional regression model is competitive across a range of situations, likely because it is more flexible than other existing functional regression models, while not subject to the curse of dimensionality. The CAM approach conveys in a compact and interpretable way the influence of the graph of the predictor trajectories on the outcome.

ACKNOWLEDGEMENTS

We wish to thank two reviewers and an associate editor for most helpful comments. In addition to several grants from 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 in Fall 2010 within the framework of the program on Analysis of Object Data.

APPENDIX

Identifiability. Consider the unconstrained continuously additive model $E(Y \mid X) = \int_{\mathcal{T}} f\{t, X(t)\}\, dt$. Here f is not identifiable. If the null space $N(K)$ of the auto-covariance operator of predictor processes X satisfies $N(K) = \{0\}$, then $\int_{\mathcal{T}} f\{t, X(t)\}dt = 0$ 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 $\int_{\mathcal{T}} f^*(t)dt = 0$.

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 $\int_{\mathcal{T}} \beta_0^*(t)dt = \beta_0|\mathcal{T}|$, while the slope function $\beta(t)$ that is of primary interest is uniquely defined when $N(K) = \{0\}$. 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 $\beta_j(t)$, $j \geq 1$, is uniquely defined. We may 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 and also for the functional linear model. The normalized model can be implemented by first obtaining an unconstrained f and then standardizing $g\{t, X(t)\} = f\{t, X(t)\} - E[f\{t, X(t)\}]$, 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.

Proof of Theorem 1. Writing $Q = Z(Z^\top Z)^{-1/2}U$, it is easy to obtain the explicit solution $\hat{\gamma} = (Z^\top Z + \lambda P)^{-1}Z^\top Y$, for $\hat{\gamma}$ as in (15), where $Y = (Y_1, \ldots, Y_n)^\top$, and

$$\begin{aligned} \hat{\theta} &= Z(Z^\top Z)^{-1/2}\{I + \lambda(Z^\top Z)^{-1/2}P(Z^\top Z)^{-1/2}\}^{-1}(Z^\top Z)^{-1/2}Z^\top Y \\ &= Q(I + \lambda D)^{-1}Q^\top Y. \end{aligned}$$

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

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

which leads to

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

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

$$\begin{aligned} E(\hat{\theta}-\hat{\theta}_u) &= Q\{(I+\lambda D)^{-1} - (I+\lambda D)^{-1}(I+\lambda D)\}Q^\top\theta \\ &= -\lambda Q(I+\lambda D)^{-1}DQ^\top\theta. \end{aligned}$$

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

$$\begin{aligned} \frac{1}{n}\|E(\hat{\theta}-\hat{\theta}_u)\|^2 &= O_p\left[\frac{\lambda^2}{n}\mathrm{tr}\{D(I+\lambda D)^{-1}Q^\top Q(I+\lambda D)^{-1}D\}\right] \\ &= O_p\left\{\frac{\lambda^2}{n}\sum_{\ell=1}^{pq}\frac{d_\ell^2}{(1+\lambda d_\ell)^2}\right\}. \end{aligned}$$

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

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

Using (A.1), the approximation error is $\|\theta_{p,q}-\theta\|_\infty = O(1/p+1/q)$ from (13) and

$$\frac{1}{n}\|E\hat{\theta}_u-\theta\|^2 = O_p\left\{\frac{1}{npq}\mathrm{tr}(I-QQ^\top)\right\} = O_p\left(\frac{n-pq}{npq}\right) = O_p\left(\frac{1}{pq}\right).$$

Proof of Theorem 2. The explicit solution is $\hat{\theta}(x) = Z_x^\top(Z^\top Z/n + \lambda P/n)^{-1}(Z^\top Y/n)$. For $R = Z^\top Z/n$, the maximum eigenvalue of $\lambda R^{-1}P/n$ is bounded by $c\lambda/(n\rho_1) = o_p(1)$ for some constant c. Applying a Taylor expansion at $\lambda = 0$, for some $\xi \in [0,\lambda]$,

$$\begin{aligned} \hat{\theta}(x) &= Z_x^\top\left\{I - \frac{\lambda}{n}R^{-1}P + \left(\frac{\xi}{n}R^{-1}P\right)^2\right\}R^{-1}\frac{1}{n}Z^\top Y \\ &= \hat{\theta}_u(x) - \frac{\lambda}{n}R^{-1}PR^{-1}\frac{1}{n}Z^\top Y + r_n, \end{aligned} \tag{21}$$

where $\hat{\theta}_u(x) = n^{-1}Z_x^\top R^{-1}Z^\top Y$ is the non-penalized version and $r_n = Z_x^\top(\xi R^{-1}P)^2/n^3 R^{-1}Z^\top Y$ is the remainder term. Since $\theta_{p,q} = Z\gamma$ and $\|\theta - \theta_{p,q}\|_\infty = O(1/p + 1/q)$, implying $\theta_{p,q} - \theta = O_p\{\theta_{p,q}(1/p + 1/q)\}$,

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

Analogously,

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

For the approximation bias, denoting $\theta_{p,q}(x) = n^{-1}Z_x^\top R^{-1}Z^\top\theta_{p,q} = Z_x^\top\gamma$ and noting that $|\theta_{p,q}(x) - \theta(x)| = O_p(1/p + 1/q)$ from (13) and $\theta_{p,q} - \theta = O_p\{\theta_{p,q}(1/p + 1/q)\}$ from above,

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

For the asymptotic variance, with a Taylor expansion similar to (21),

$$\begin{aligned} \mathrm{var}\{\hat{\theta}(x)\} &= \frac{\sigma^2}{n}Z_x^\top\left\{I - \frac{\lambda}{n}R^{-1}P + \left(\frac{\xi}{n}R^{-1}P\right)^2\right\}^2 R^{-1}Z_x \\ &= \frac{\sigma^2}{n}\left[Z_x^\top R^{-1}Z_x - \frac{2\lambda}{n}Z_x^\top R^{-1}PR^{-1}Z_x\{1 + o_p(1)\}\right]. \end{aligned} \tag{22}$$

As the maximal eigenvalue of $\lambda R^{-1}/n$ is bounded from above by $\lambda/(n\rho_1) = o_p(1)$, the second term in (22) reflects a reduction of variance that corresponds to a higher order term. Asymptotically, the leading term of the variance is $v(x) = \sigma^2 Z_x^\top R^{-1}Z_x/n$. As for an arbitrary X, $Z_X^\top Z_X = \sum_{\ell=1}^{pq} Z_{X,\ell}^2 \le \sum_{\ell=1}^{pq} Z_{X,\ell} = 1$, the maximal eigenvalue of $R = n^{-1}\sum_{i=1}^n Z_{X_i}Z_{X_i}^\top$ is not greater than $\mathrm{tr}(R) \le 1$, implying that $v(x)$ is bounded in probability by $\sigma^2 n^{-1} \le v(x) \le \sigma^2(n\rho_1)^{-1}$.

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 $\min(p^2, q^2)/n \to \infty$. The shrinkage bias needs to satisfy $b_\lambda(x)/\{v(x)\}^{1/2} = O_p(1)$, which is guaranteed by $\lambda^2/(n\rho_1^2) = O_p(1)$. 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 $\left[\hat{\theta}(x) - E\{\hat{\theta}(x)\}\right] \Big/ [\mathrm{var}\{\hat{\theta}(x)\}]^{1/2}$ converges to a standard normal distribution. Writing

$\hat{\theta}(x) - E\{\hat{\theta}(x)\} = n^{-1} Z_x (R + n^{-1}\lambda P)^{-1} Z^\top (Y - \theta) = \sum_{i=1}^n a_i \epsilon_i$, 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 verify that $\max_{1 \le i \le n} a_i^2 = o_p(\sum_{i=1}^n a_i^2) = o_p[\mathrm{var}\{\hat{\theta}(x)\}]$.

As the maximal eigenvalue of $Z_x Z_x^\top$ is not greater than $Z_x^\top Z_x \le 1$ for any x, $\mathrm{tr}(AB) \le \rho_A \mathrm{tr}(B)$ for nonnegative definite matrices A and B, where ρ_A is the maximal eigenvalue of A, implies

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

Applying a similar Taylor expansion as above and observing $\lambda/(n\rho_1) = o_p(1)$, the above quantity is bounded in probability by $(n\rho_1)^{-2}\mathrm{tr}(Z_i Z_i^\top) \le (n\rho_1)^{-2}$. Then $\lambda^2/(n\rho_1^2) = O_p(1)$ and $\lambda \to \infty$ imply that $\max_{1 \le i \le n} a_i^2 / v(x) \le 1/(n\rho_1^2)$ converges to 0 in probability, completing the proof.

References

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

Cardot, H., Crambes, C., Kneip, A. & Sarda, P. (2007). Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics and Data Analysis* **51**, 4832–4848.

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

Carroll, R., Aity, A., Mammen, E. & Yu, K. (2008). Nonparametric additive regression for repeatedly measured data. *Biometrika* **36**, 383–398.

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

Escabias, M., Aguilera, A. M. & Valderrama, M. J. (2004). Principal component estimation of functional logistic regression discussion of two different approaches. *Journal of Nonparametric Statistics* **16**, 365–384.

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

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

Friedman, J. & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association* **76**, 817–823.

Gasser, T., Müller, H.-G., Köhler, W., Molinari, L. & Prader, A. (1984). Nonparametric regression analysis of growth curves. *Annals of Statistics* **12**, 210–229.

Goutis, C. (1998). Second-derivative functional regression with applications to near infra-red spectroscopy. *Journal of the Royal Statistical Society: Series B* **60**, 103–114.

Hall, P., Müller, H.-G. & Yao, F. (2009). Estimation of functional derivatives. *Annals of Statistics* **37**, 3307–3329.

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. *Journal of the Royal Statistical Society: Series B* **64**, 411–432.

Kneip, A. & Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *Annals of Statistics* **20**, 1266–1305.

Mammen, E. & Park, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *Annals of Statistics* **33**, 1260–1294.

Marx, B. & Eilers, B. (1996). Flexible smoothong with B-splines and penalties (with comments and rejoinder). *Statistical Science* **11**, 89–121.

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

Müller, H.-G. & Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103**, 1534–1544.

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

Ravikumar, P., Lafferty, J., Liu, H. & Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B* **71**, 1009–1030.

Reiss, P. & Ogden, R. (2010). Functional generalized linear models with images as predictors. *Biometrics* **66**, 61–69.

Song, J., Deng, W., Lee, H. & Kwon, D. (2008). Optimal classification for time-course gene expression data using functional data analysis. *Computational Biology and Chemistry* **32**, 426–432.

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. *Molecular Biology of the Cell* **9**, 3273–3297.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics* **13**, 689–705.

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

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.

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. *Annals of Statistics* **36**, 228–260.