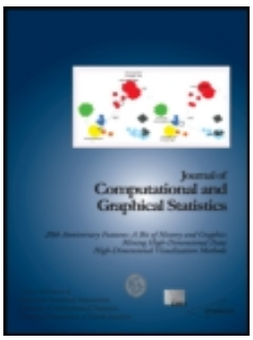

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Functional Generalized Additive Models

Mathew W. McLean, Giles Hooker, Ana-Maria Staicu, Fabian Scheipl & David Ruppert

To cite this article: Mathew W. McLean, Giles Hooker, Ana-Maria Staicu, Fabian Scheipl & David Ruppert (2014) Functional Generalized Additive Models, Journal of Computational and Graphical Statistics, 23:1, 249-269, DOI: 10.1080/10618600.2012.729985

To link to this article: https://doi.org/10.1080/10618600.2012.729985

Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

Functional Generalized Additive Models

Mathew W. McLean, Giles Hooker, Ana-Maria Staicu, Fabian Scheipl, and David Ruppert

We introduce the functional generalized additive model (FGAM), a novel regression model for association studies between a scalar response and a functional predictor. We model the link-transformed mean response as the integral with respect to t of $F\{X(t), t\}$ where $F(\cdot, \cdot)$ is an unknown regression function and $X(t)$ is a functional covariate. Rather than having an additive model in a finite number of principal components as by Müller and Yao (2008), our model incorporates the functional predictor directly and thus our model can be viewed as the natural functional extension of generalized additive models. We estimate $F(\cdot, \cdot)$ using tensor-product B-splines with roughness penalties. A pointwise quantile transformation of the functional predictor is also considered to ensure each tensor-product B-spline has observed data on its support. The methods are evaluated using simulated data and their predictive performance is compared with other competing scalar-on-function regression alternatives. We illustrate the usefulness of our approach through an application to brain tractography, where $X(t)$ is a signal from diffusion tensor imaging at position, t, along a tract in the brain. In one example, the response is disease-status (case or control) and in a second example, it is the score on a cognitive test. The FGAM is implemented in R in the `refund` package. There are additional supplementary materials available online.

Key Words: Diffusion tensor imaging; Functional data analysis; Functional regression; P-spline.

1. INTRODUCTION

This article studies regression with a functional predictor and a scalar response. Suppose one observes data $\{(X_i(t), Y_i) : t \in \mathcal{T}\}$ for $i = 1, \ldots, N$, where X_i is a real-valued, continuous, square-integrable, random curve on the compact interval $\mathcal{T}$ and Y_i is a scalar. We assume that the predictor, $X(\cdot)$, is observed at a dense grid of points. The problem

Mathew W. McLean is Research Assistant Professor at the Institute for Applied Mathematics and Computational Science, Texas A&M University, College Station, TX 77843 (E-mail: *mmclean@stat.tamu.edu*). Giles Hooker is Associate Professor at the Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853 (E-mail: *giles.hooker@cornell.edu*). Ana-Maria Staicu is Assistant Professor at the Department of Statistics, North Carolina State University, Raleigh, NC 27695 (E-mail: *staicu@stat.ncsu.edu*). Fabian Scheipl is postdoc candidate at the Department of Statistics, Ludwig Maximilian University of Munich, Munich 80333 (E-mail: *fabian.scheipl@stat.uni-muenchen.de*). David Ruppert is Jr. Professor of Engineering, School of Operations Research and Information Engineering, and Professor of Statistical Science, Department of Statistical Science, Cornell University, 1170 Comstock Hall, Ithaca, NY 14853 (E-mail: *dr24@cornell.edu*).

Journal of Computational and Graphical Statistics, Volume 23, Number 1, Pages 249–269
DOI: 10.1080/10618600.2012.729985
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jcgs*.

addressed here is estimation of $E(Y_i|X_i)$, which is assumed independent of i. We introduce the model

$$g\{E(Y_i|X_i)\} = \theta_0 + \int_{\mathcal{T}} F\{X_i(t), t\}\, dt, \tag{1}$$

where θ_0 is the intercept, g is a known link function, and F is an unspecified smooth function to be estimated. As a special case, when $g(x) = x$ and $F(x, t) = \beta(t)X_i(t)$, we obtain the most commonly used regression model in functional data analysis, the functional linear model (Ramsay and Dalzell 1991), henceforth the FLM,

$$E(Y_i|X_i) = \theta_0 + \int_{\mathcal{T}} \beta(t)X_i(t)\, dt, \tag{2}$$

where $\beta(\cdot)$ is the functional coefficient with $\beta(t)$ describing the effect on the response of the functional predictor at time t. The FLM can be thought of as multiple linear regression with an infinite number of predictors, as we now explain. Let $t_{ij} = t_j$ for $1, \ldots, J$ denote the observation times for the curves $X_i(\cdot)$; then the usual multiple linear regression model $E(Y_i|X_i(t_1), \ldots, X_i(t_J)) = \beta_0 + \sum_{i=1}^{J} \beta_j X_i(t_j) = \beta_0 + J^{-1} \sum_{i=1}^{J} \beta_j' X_i(t_j)$ can be viewed as a Riemann sum approximation that converges to Equation (2) as $J \to \infty$. This model has been extended to a functional generalized linear model (GLM), that is, a model of the form $g\{E(Y_i|X_i)\} = \theta_0 + \int_{\mathcal{T}} X_i(t)\beta(t)dt$ (e.g., James 2002; Müller and Stadtmüller 2005).

Now consider an additive model of the form $E\{Y_i|X_i(t_1), \ldots, X_i(t_J)\} = \theta_0 + \sum_{j=1}^{J} f_j\{X_i(t_j)\}$, where the f_j's are unspecified smooth functions. The basic idea is to rewrite the model as $E\{Y_i|X_i(t_1), \ldots, X_i(t_J)\} = \theta_0 + J^{-1} \sum_{j=1}^{J} F\{X_i(t_j), t_j\}$, and then let $J \to \infty$ and add a link function. The model obtained is our model (1). We call model (1) the functional generalized additive model (FGAM). Our modeling approach provides greater flexibility, as it does not make the strong assumption of linearity between the functional predictor and the functional parameter. To overcome the so-called curse of dimensionality, we will perform smoothing in both the x and t components of $F(\cdot, \cdot)$. Just as the FLM is the natural extension of linear models to functional data, our model is the natural extension of generalized additive models (GAMs) to functional data.

There are few instances in the literature of nonparametric, additive structures being used for scalar on function regression models. James and Silverman (2005) and Müller and Yao (2008) both considered GAMs that use linear functionals of the predictor curves as covariates. The former approach regresses on a finite number of functional principal components scores and the latter approach searches for linear functionals using projection pursuit. Both models rely strongly on the linear directions they estimate; in contrast, our modeling approach regresses on the functional predictors directly. A model that is additive in the principal component scores is not additive in $X(t)$ itself, and vice versa. Therefore, our FGAM and the additive model by Müller and Yao (2008) are different and should be considered complementary, rather than competitors.

Additive models are attractive for a number of reasons (see, e.g., Buja, Hastie, and Tibshirani 1989, and Hastie and Tibshirani 1990, which are standard, early references). Initial work in the area advocated smoothing splines and backfitting for fitting these models. Additional theory for backfitting was developed in a series of articles by Opsomer and Ruppert (1997, 1998, 1999). An alternative to classical backfitting is the smooth backfitting by Mammen, Linton, and Nielsen (1999). Though not as widely used, it has been shown

to offer a number of advantages (see, e.g., Nielsen and Sperlich 2005, for a discussion of its implementation and practical performance). Recently, penalized regression splines have proven successful in a number of applications. Estimation in this case is most often done with penalized, iteratively reweighted least squares (P-IRLS) with smoothing parameters chosen using generalized cross-validation (GCV; see, e.g., Marx and Eilers 1998; Ruppert, Wand, and Carroll 2003; Wood 2006b). This is the approach we adopt to estimate the FGAM. In general, additive models offer increased flexibility and potentially lower estimation bias than linear models while having less variance in estimation and being less susceptible to the curse of dimensionality than models that make no additivity assumptions. The proposed model (1) provides greater flexibility than the FLM, while still facilitating interpretation and estimation.

One area where this increased flexibility is useful is diffusion tensor imaging (DTI), which we consider in Section 5. The dataset contains closely spaced evaluations of measures of neural functioning on multiple tracts in the brain for patients with multiple sclerosis and healthy controls. We will use these measurements as regressors and predict multiple health outcomes to gain a better understanding of how the disease is related to DTI signals. Our model is able to quantify the effect that the functional predictor has on the response at each position along the tract, something that a model such as the GAM by Müller and Yao (2008) is unable to do, since it uses principal component scores and hence loses information about tract location. Another potential application of FGAM is to study how a risk factor trajectory such as body mass index or systolic blood pressure is related to a health outcome such as developing hypertension (e.g., see the study by Li, Wang, and Wang 2007). Our FGAM can locate times of life when the risk factor has its greatest effect; this is not possible if principal component scores are used in a GAM.

For estimation of the model (1), we will use P-splines (Eilers and Marx 1996). However, there will be some differences from standard fitting of tensor product P-splines. Namely, our design matrix is obtained from integrating products of B-splines over functional covariates. P-splines offer many computational advantages. Additional scalar or functional predictors can be incorporated in a simple way and will not require backfitting. Both types of predictors can be included in either a linear or an additive fashion. Though we use P-splines, our estimation procedure can incorporate other bases and penalties for some or all of the covariates. We use well-developed, efficient techniques for the computations (Wood 2006b; Ramsay, Hooker, and Graves 2009).

We also propose transforming the functional predictors using the empirical cumulative distribution function (empirical cdf) at fixed values of t. This transformation is convenient for estimation purposes and retains the interpretation advantages provided by the FGAM when the raw curves are used. Considering again the DTI data example, when this transformation is used, we can now infer the effect on the response of a subject being in the pth-quantile for the functional predictor at a particular location along the tract.

To see how our model can aid in uncovering the underlying structure of a functional regression problem, consider Figure 1. The figure shows the estimated surface, $\widehat{F}(\cdot, \cdot)$, for one of the functional predictors in the DTI dataset when the response is disease status ($= 1$ if the subject has the disease). Overlaid on the surface are the observed functional predictor values for two subjects. The surface is nonlinear in x, so an FLM based on the predictors may be inadequate for this problem. We see that for the most part, the solid

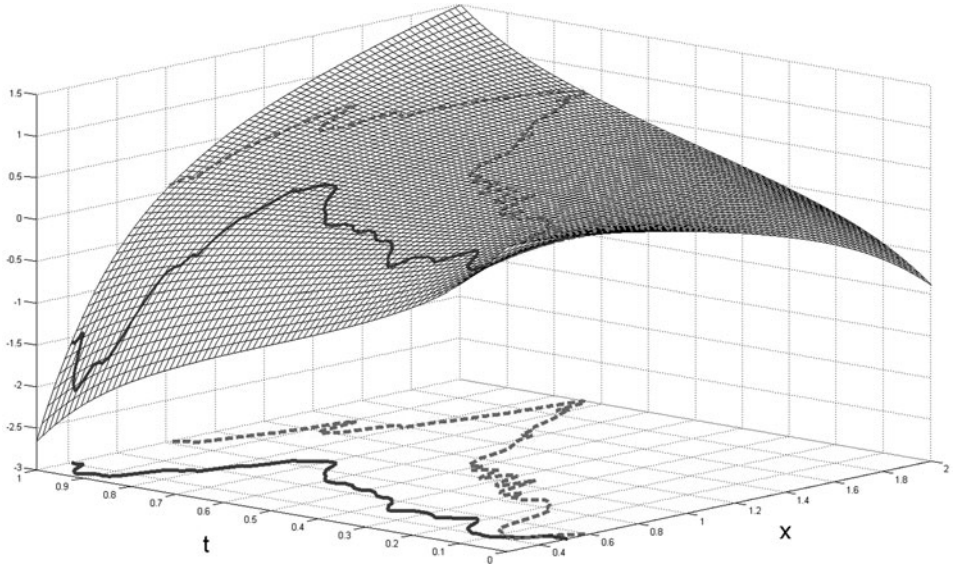

Figure 1. Estimated surface $\hat{F}(x, t)$ and two predictor curves for the DTI dataset. The solid curve belongs to a control and the dashed curve belongs to an MS patient.

curve, belonging to a control subject, takes smaller values on the surface than the dashed curve, which belongs to a MS patient, does; thus, the subject with MS will have a higher fitted value and is more likely to be classified as having the disease. It will be shown for this dataset that the added generality of our approach leads to improved predictive accuracy over the FLM.

We also show how standard error estimates for the parameters of the FGAM are obtained and examine the performance of confidence bands constructed from these standard errors through a simulation study. These are used to make approximate inferences about the estimated surface and the estimated second derivative surface $\partial^2/\partial x^2 \widehat{F}(x, t)$ that can be used to detect nonlinearity in x. Several diagnostic plots such as the one in Figure 1 are available for exploring the relationship between the predictors and the response.

The article is organized as follows. Section 2 introduces the FGAM in more detail. Our estimation procedure using P-splines is discussed in Section 3. Section 4 applies our model to simulated datasets and compares it with some standard regression models used in functional data analysis. Section 5 discusses the results of applying our model to the DTI dataset. Section 6 concludes with a brief discussion and mentions some possible extensions.

2. FUNCTIONAL GENERALIZED ADDITIVE MODEL

In this section, we introduce our representation for $F(\cdot, \cdot)$, describe the identifiability constraints, and discuss a transformation of the functional predictor. It is assumed that $\mathcal{T} = [0, 1]$ and that $X(\cdot)$ takes values in a bounded interval that, without loss of generality, can be taken as $[0, 1]$. The latter assumption is guaranteed by the proposed transformation of the functional predictors discussed in Section 2.2.

We will model $F(\cdot, \cdot)$ using tensor products of B-splines. Splines are commonly used for estimation of functional linear models. For example, smoothing splines were used by Crambes, Kneip, and Sarda (2009) and Yuan and Cai (2010) and penalized splines were considered by Cardot, Ferraty, and Sarda (2003) and Goldsmith et al. (2011). These articles impose smoothness using a penalty on the integrated, squared second derivative of the coefficient function. Instead, we use the popular P-splines by Eilers and Marx (1996). P-splines use low rank B-splines bases with equally spaced knots and a simple difference penalty on adjacent coefficients to control smoothness.

The many advantages of using P-spline estimators in additive modeling were discussed in detail by Marx and Eilers (1998). The implementation with P-splines will make it possible to estimate all the components of the model at once. While backfitting could be implemented for the case of multiple predictors, it is not feasible for estimating Equation (1). It will be shown that the fitted values for the FGAM are linear in the tensor product B-spline coefficients so we actually have a penalized GLM. By using fewer knots than there are observations, the size of the system of equations for the estimation is reduced. Penalized splines are fairly insensitive to the position and the number of knots compared to unpenalized splines. Also, unlike smoothing splines, P-splines allow any degree of B-spline to be used with any order of differencing for the penalty.

2.1 NOTATION AND IDENTIFIABILITY CONSTRAINTS

A bivariate spline model is used for $F(\cdot, \cdot)$ so that

$$F(x,t) = \sum_{j=1}^{K_x} \sum_{k=1}^{K_t} \theta_{j,k} B_j^X(x) B_k^T(t), \tag{3}$$

where $\{B_j^X(x) : j = 1, \ldots, K_x\}$ and $\{B_k^T(x) : k = 1, \ldots, K_t\}$ are spline bases on $[0, 1]$. We will use B-spline bases. It follows from combining Equations (1) and (3), that we obtain the GLM

$$g\{E(Y_i|X_i)\} = \theta_0 + \int_{\mathcal{T}} F\{X_i(t), t\}dt = \theta_0 + \sum_{j=1}^{K_x} \sum_{k=1}^{K_t} \theta_{j,k} Z_{j,k}(i), \tag{4}$$

where $Z_{j,k}(i) = \int_{\mathcal{T}} B_j^X\{X_i(t)\}B_k^T(t)dt$. Each $Z_{j,k}(i)$ can be approximated by, say, Simpson's rule.

For identifiability, we require $\sum_{i=1}^{N} \int_{\mathcal{T}} F(X_i(t), t)\, dt = 0$ (Wood 2006b, Sec. 4.2). This constraint may not be enough to ensure identifiability on its own, however, so we must perform a further check for numerical rank deficiency during fitting. The details are explained in the next section. Additional discussion of identifiability for the FGAM is also provided in Appendix A available in the online supplementary materials.

2.2 TRANSFORMATION OF THE PREDICTORS

Depending on the number of B-splines used for each axis, there could be a particular tensor product of B-splines that has no observed data on its support. This would lead to $Z_{j,k}(i) = 0$ for all i for some j, k pair, resulting in the design matrix containing a column of zeros. One remedy for this is to transform $X(t)$ by $G_t(x) := P\{X(t) < x\}$ for each value

of t. Our model becomes

$$g\{E(Y_i|X_i)\} = \theta_0 + \int_{\mathcal{T}} F[G_t\{X_i(t)\}, t]dt = \theta_0 + \sum_{j=1}^{K_x}\sum_{k=1}^{K_t}\theta_{j,k}\int_{\mathcal{T}} B_j^G[G_t\{X_i(t)\}]B_k^T(t)dt, \tag{5}$$

where $B^G(\cdot)$ is a new B-spline basis with support on $[0, 1]$. Loosely, the data are being "stretched out" to fill the entire space that the grid of B-splines will cover. For any t, the transformed points will lie uniformly between $[0, 1]$. Though the estimation procedure is the same in both cases, clearly, $F(\cdot, \cdot)$ in Equation (5) will have a different estimate from $F(\cdot, \cdot)$ in Equation (1). We estimate $G_t(\cdot)$ using the empirical cdf $\widehat{G}_t(x) = n^{-1}\sum_{i=1}^n I\{X_i(t) < x\}$, where $I\{A\} = 1$ if condition A is true and $I\{A\} = 0$ otherwise. Once the $Z_{j,k}(i)$'s have been estimated, the fitting procedure is analogous to the case when the cdf transformation is not used. Another advantage of using this approach is that it does not require any assumptions about the range of the predictors. Besides the computational advantages, this transformation retains the benefit of ease of interpretation. In fact, $F(p, t)$ is the effect of $X(t)$ being at its pth quantile.

Another potentially useful transformation we do not pursue in this article is $\hat{H}_t(x) = n^{-1}\sum_{i=1}^n \Phi[\frac{x-X_i(t)}{h_t}]$, where $\Phi(\cdot)$ denotes the standard normal cdf and h_t is a user chosen bandwith that can depend on t. The advantage of this transformation over the empirical cdf transformation is that future observations falling below (above) the minimum (maximum) value of the training data at a particular t are not all assigned the value zero (one).

Due to the penalization used later when fitting the FGAM, parameter estimates can still be obtained when the design matrix has a column of zeros. However, we expect our transformation will improve both the numerical and statistical stability of our estimates. Note also that if there exists any pointwise transformation, $H_t(\cdot)$, such that $g\{E(Y_i|X_i)\} = \int_{\mathcal{T}} \beta(t)H_t\{X_i(t)\}\,dt$, then the FGAM will still hold; and similarly, for any model of the form (5) for a general transformation $G_t(\cdot)$. Thus, the FGAM is invariant to transformations of the predictor, unlike the FLM.

3. ESTIMATION

In this section, we present the estimation procedure for $F(\cdot, \cdot)$. First, we review P-spline type penalties and discuss penalized GLMs and the selection of smoothing parameters. We then describe the estimated surface and discuss construction of pointwise confidence bands for these estimates. We conclude the section by showing how to include additional functional and nonfunctional predictors in the model.

3.1 ROUGHNESS PENALTIES

Smoothing can be achieved by using row and column penalties as in the article by Marx and Eilers (1998). The row penalty is $\lambda_1\sum_{j=d+1}^{K_x}(\Delta_j^d\theta_{j,k})^2$, where $\Delta_j^d\theta_{j,k}$ is the dth difference of the sequence $\theta_{j-d,k}, \ldots, \theta_{j,k}$ (k held fixed). The column penalty is $\lambda_2\sum_{k=d+1}^{K_t}(\Delta_k^d\theta_{j,k})^2$, where $\Delta_k^d\theta_{j,k}$ is the dth difference of the sequence $\theta_{j,k-d}, \ldots, \theta_{j,k}$ (j held fixed). Selection of the penalty parameters λ_1 and λ_2 is discussed in Section 3.2.

Proceeding similarly to Marx and Eilers (2005), we first place the $Z_{j,k}(i)$'s in a matrix as follows. Let $\mathbf{Z}_i = \text{vec}\{\mathbb{Z}(i)$ be the $\kappa_x K_t$-vector obtained by stacking the columns of $\mathbb{Z}(i) = [Z_{j,k}(i)]_{j=1,\ldots,K_x}^{k=1,\ldots,K_t}$, and let $\mathbb{Z} = [\mathbf{Z}_1\ \mathbf{Z}_2 \ldots \mathbf{Z}_N]^T$. The penalty matrix is given by

$$\mathbb{P} = \lambda_1 \mathbb{P}_1^T \mathbb{P}_1 + \lambda_2 \mathbb{P}_2^T \mathbb{P}_2, \tag{6}$$

with $\mathbb{P}_1 = \mathbb{D}_x \otimes \mathbb{I}_{K_t}$, $\mathbb{P}_2 = \mathbb{I}_{K_x} \otimes \mathbb{D}_t$ where $\mathbb{I}_p$ is the $p \times p$ identity matrix, $\otimes$ is the Kronecker product, and $\mathbb{D}_x$ and $\mathbb{D}_t$ are matrix representations of the row and column difference penalties with dimensions $(K_x - d_x) \times K_x$ and $(K_t - d_t) \times K_t$, respectively. The parameter, d, denotes the prespecified degree of differencing. Note that additional penalties such as an overall ridge penalty could also be incorporated.

To incorporate the intercept, a leading column of ones must be added to $\mathbb{Z}$ and a leading column of zeros must be added to $\mathbb{P}_1$ and $\mathbb{P}_2$. Throughout the rest of the article, this has been done unless otherwise indicated. When we do not wish to consider the intercept, $\mathbb{M}_{[-i,-j]}$ will denote the matrix $\mathbb{M}$ with its ith row and jth column removed and $\mathbf{v}_{[-i]}$ will denote the vector $\mathbf{v}$ excluding its ith entry.

3.2 PENALIZED GLMS AND SMOOTHING PARAMETER SELECTION

Let the response vector, $\mathbf{Y}$, be from an exponential family with density having the form $f_Y(\mathbf{y}; \boldsymbol{\zeta}, \phi) = \prod_{i=1}^N \exp[\{y_i\zeta_i - b(\zeta_i)\}/a(\phi) + c(y_i, \phi)]$, where $\boldsymbol{\zeta}$ is the canonical parameter vector with components satisfying $\zeta_i = (b')^{-1}(\mu_i)$ and ϕ is the dispersion parameter. Parameterizing $E(\mathbf{Y}|\mathbb{X})$ as a standard GLM with known link function, $g(\cdot)$, let $\boldsymbol{\eta} := \mathbb{Z}\boldsymbol{\theta}$ and $\boldsymbol{\mu} := E(\mathbf{Y}|\mathbb{X})$, so that $\boldsymbol{\eta} = g(\boldsymbol{\mu})$. The constraint discussed in Section 2.1 is enforced by requiring $\mathbf{1}^T\mathbb{Z}_{[,-1]}\boldsymbol{\theta} = 0$. Formally, this is done by obtaining the QR decomposition of $(\mathbf{1}^T\mathbb{Z}_{[,-1]})^T = [\mathbf{Q}_1\ \mathbb{Q}_2][\begin{smallmatrix}\mathbb{R}_1\\ \mathbb{R}_2\end{smallmatrix}]$, where $\mathbb{Q}_2$ has dimension $(1 + K_xK_t) \times K_xK_t$. The constrained optimization problem is then replaced by an unconstrained optimization (outlined below) over $\boldsymbol{\theta}_q$, where $\boldsymbol{\theta}_q$ is such that $\boldsymbol{\theta} = \mathbb{Q}_2\boldsymbol{\theta}_q$. For notational simplicity, for any matrix $\mathbb{M}$, define $\tilde{\mathbb{M}} = \mathbb{M}\mathbb{Q}_2$.

The penalized log-likelihood to be maximized is

$$l(\boldsymbol{\theta}_q; \lambda_1, \lambda_2) = \sum_{i=1}^N \log\{f_Y(y_i; \zeta_i, \phi)\} - \lambda_1||\tilde{\mathbb{P}}_1\boldsymbol{\theta}_q||^2 - \lambda_2||\tilde{\mathbb{P}}_2\boldsymbol{\theta}_q||^2.$$

The coefficients are estimated using P-IRLS. Specifically, at the $(m+1)$th iteration we take

$$\widehat{\boldsymbol{\theta}}_{q,m+1} = \left(\tilde{\mathbb{Z}}^T\widehat{\mathbb{W}}_m\tilde{\mathbb{Z}} + \lambda_1\tilde{\mathbb{P}}_1^T\tilde{\mathbb{P}}_1 + \lambda_2\tilde{\mathbb{P}}_2^T\tilde{\mathbb{P}}_2\right)^{-1}\tilde{\mathbb{Z}}^T\widehat{\mathbb{W}}_m\widehat{\mathbf{u}}_m, \tag{7}$$

where $\widehat{\mathbf{u}}_m$ is the current estimate of the adjusted dependent variable vector, $\mathbf{u}$, and $\widehat{\mathbb{W}}_m$ is the current estimate of the diagonal weight matrix, $\mathbb{W}$. The components of $\mathbf{u}$ are given by $u_i = \eta_i + (y_i - \mu_i)g'(\mu_i)$. The ith diagonal element of $\mathbb{W}$ is $w_{ii} = 1/\{V(\mu_i)[g'(\mu_i)]^2\}$, with $V(\mu_i) = b''(\zeta_i)$. To initialize the algorithm, use $\boldsymbol{\mu}_0 = \mathbf{Y}$ and $\boldsymbol{\eta}_0 = g(\mathbf{Y})$, adjusting y_i if necessary to avoid $\eta_i = \infty$.

To efficiently construct Equation (7) and to detect rank deficiency, the following procedure is used. First, use the QR-decomposition to form $\mathbb{W}^{1/2}\tilde{\mathbb{Z}} = \mathbb{Q}\mathbb{R}$, where $\mathbb{Q}$ is orthogonal, $\mathbb{R}$ is upper triangular, and $\mathbb{W}^{1/2} = \text{diag}(w_{11}^{1/2}, \ldots, w_{NN}^{1/2})$. Next, use the Choleski decomposition to obtain $\mathbb{Q}_2^T\mathbb{P}\mathbb{Q}_2 = \mathbb{L}^T\mathbb{L}$. Pivoting should be used here because $\mathbb{P}$ is

positive semidefinite instead of positive definite. Now, from a singular value decomposition form $[\mathbb{R}^T \ \mathbb{L}^T]^T = \mathbb{U}\mathbb{D}\mathbb{V}^T$, where $\mathbb{U}$ and $\mathbb{V}$ are orthogonal and $\mathbb{D}$ is a diagonal matrix containing the singular values. At this point, we ensure identifiability by removing the columns and rows of $\mathbb{D}$ and the columns of $\mathbb{U}$ and $\mathbb{V}$ corresponding to singular values that are less than the square root of the machine precision times the largest singular value (Wood 2006b, p. 183). It then follows that Equation (7) can be obtained from $\widehat{\boldsymbol{\theta}}_{q,m+1} = \mathbb{V}\mathbb{D}^{-1}\mathbb{U}_1^T\mathbb{Q}^T\mathbb{W}^{1/2}\widehat{\mathbf{u}}_m$, where $\mathbb{U}_1$ is the submatrix of $\mathbb{U}$ satisfying $\mathbb{R} = \mathbb{U}_1\mathbb{D}\mathbb{V}^T$. At the final iteration, say M, our solution for $\boldsymbol{\theta}$ is given by $\widehat{\boldsymbol{\theta}} = \mathbb{Q}_2\widehat{\boldsymbol{\theta}}_{q,M}$ and it can be shown that this satisfies $\mathbf{1}^T\mathbb{Z}_{[,-1]}\widehat{\boldsymbol{\theta}} = 0$ as required (Wood 2006b, Sec. 1.8.1).

Generalized cross-validation can be used to choose the smoothing parameters (see Wood 2004, Sec. 4.5.4, for justification of its use for nonidentity link GAMs). The GCV score for λ_1 and λ_2 is given by

$$\text{GCV}(\lambda_1, \lambda_2) = \frac{nD(\mathbf{Y}; \widehat{\boldsymbol{\mu}} : \lambda_1, \lambda_2)}{\{n - \gamma \text{tr}(\mathbb{H})\}^2}, \tag{8}$$

where $\mathbb{H}$ is known as the influence matrix and is related to the fitted values by $\widehat{\boldsymbol{\mu}} := g^{-1}(\mathbb{Z}\widehat{\boldsymbol{\theta}}_M) = g^{-1}(\mathbb{H}\mathbf{u}_M)$ and $D(\mathbf{Y}; \widehat{\boldsymbol{\mu}} : \lambda_1, \lambda_2)$ denotes the model deviance. The model deviance is defined to be twice the difference between the log-likelihoods of the saturated model, which has one parameter for each observation, and the given model. Formulas for the deviance for some common GLMs are given by McCullagh and Nelder (1989, Sec. 2.3); for example, for an identity link GLM, $D(\mathbf{Y}; \widehat{\boldsymbol{\mu}} : \lambda_1, \lambda_2) = ||\mathbf{Y} - \mathbb{H}\mathbf{Y}||^2$. The constant $\gamma \geq 1$ is usually chosen to take values between 1.2 and 1.4 to combat the tendency of GCV to undersmooth. For additional safeguards against undersmoothing, lower bounds could also be placed on the smoothing parameters.

A choice must be made on the order in which the P-IRLS and the smoothing parameter selection iterations are performed. For what is termed outer iteration, for each pair of smoothing parameters considered, a GAM is estimated using P-IRLS until convergence. The other possibility, known as performance iteration, is to optimize the smoothing parameters at each iteration of the P-IRLS algorithm. The latter approach can be faster than outer iteration; however, it is more susceptible to convergence problems in the presence of multicollinearity (Wood 2006b, Chap. 4).

Our model can conveniently be fit in R using the `mgcv` package (Wood 2011, 2004). The details of how this is done are available in Appendix B of the supplementary materials. We use outer iteration and Newton's method for minimizing the GCV score, the package defaults. Using this package also allows for many possible extensions (e.g., mixed effects terms, formal model selection, alternative estimation procedures, etc.) beyond the scope of the current article. Our code is available in the online supplementary materials as well as in the R package `refund`.

3.3 Estimated Surface

For a given $\widehat{\boldsymbol{\theta}}$, we can evaluate the estimated surface at any grid of points in its domain. Let $\mathbf{X}$ be an arbitrary column vector of length n_1 taking values in the range of $X(\cdot)$ and $\mathbf{T}$ be the observation times or any vector of length n_2 taking values in $[0, 1]$. We let $\widehat{\mathbf{F}}$ denote the estimated surface evaluated on the mesh defined by $\mathbf{X}$ and $\mathbf{T}$. To obtain $\widehat{\mathbf{F}}$,

let $\mathbb{B}_x$ be the $n_1 n_2 \times K_x$ matrix of x-axis B-splines evaluated at $\mathbf{X} \otimes \mathbf{1}_{n_2}$, that is, $\mathbb{B}_x = [\mathbf{B}_1^X(\mathbf{X} \otimes \mathbf{1}_{n_2}) \ldots \mathbf{B}_{K_x}^X(\mathbf{X} \otimes \mathbf{1}_{n_2})]$, where $\mathbf{1}_n$ denotes a column vector of length n. Similarly, define $\mathbb{B}_t$ as the $n_1 n_2 \times K_t$ matrix of B-splines evaluated at $\mathbf{1}_{n_1} \otimes \mathbf{T}$. Next, define the $n_1 n_2 \times K_x K_t$ matrix

$$\mathbb{B} = \left(\mathbb{B}_x \otimes \mathbf{1}_{K_t}^T\right) \odot \left(\mathbf{1}_{K_x}^T \otimes \mathbb{B}_x\right), \tag{9}$$

where $\odot$ denotes element-wise matrix multiplication. The estimated surface is then given by $\widehat{\mathbf{F}} = \mathbb{B}\widehat{\boldsymbol{\theta}}_{[-1]}$.

3.4 STANDARD-ERROR BANDS

For a response from any exponential family distribution, one simple way to construct approximate, pointwise confidence bands for $\widehat{F}(x,t)$ conditional on the estimated smoothing parameters is to use a sandwich estimator in the same manner as Hastie and Tibshirani (1990, Sec. 6.8.2) and Marx and Eilers (1998). However, we found through our simulation studies that these intervals do not have adequate coverage for our model, a result also noticed for univariate GAMs by Wood (2006a). This is because these intervals assume $\widehat{\boldsymbol{\theta}}$ is unbiased, which will not be the case when $\boldsymbol{\theta} \neq \mathbf{0}$, due to the penalization involved in the estimation.

To overcome the bias in the parameter estimation, we use the Bayesian approach by Wahba (1983). Using the improper prior $\pi(\boldsymbol{\theta}) \propto \exp\left(-\boldsymbol{\theta}^T \mathbb{P} \boldsymbol{\theta}/2\right)$, it can be shown that

$$\boldsymbol{\theta} | \mathbb{Z}^T \mathbb{W} \mathbf{u}, \lambda_1, \lambda_2 \sim N\left((\mathbb{Z}^T \mathbb{W} \mathbb{Z} + \mathbb{P})^{-1} \mathbb{Z}^T \mathbb{W} \mathbf{u}, (\mathbb{Z}^T \mathbb{W} \mathbb{Z} + \mathbb{P})^{-1} \phi\right)$$

(see, e.g., Wood 2006b, Sec. 4.8). To estimate $\mathbb{W}$, we use the estimated weight matrix at the final P-IRLS iteration, $\widehat{\mathbb{W}}_M$. If it is necessary to estimate the dispersion parameter, ϕ, we use $\widehat{\phi} = \sum_{i=1}^n V(\hat{\mu}_i)^{-1}(y_i - \hat{\mu}_i)^2/[N - \text{tr}(\mathbb{H})]$. Letting $\mathbb{V}_{\widehat{\boldsymbol{\theta}}} = (\mathbb{Z}^T \widehat{\mathbb{W}}_M \mathbb{Z} + \mathbb{P})^{-1}\widehat{\phi}$ and recalling that the estimated surface is given by $\widehat{\mathbf{F}} = \mathbb{B}\widehat{\boldsymbol{\theta}}_{[-1]}$, where $\mathbb{B}$ is defined in Equation (9), the variance of $\widehat{\mathbf{F}}$ is given by $\text{var}\{\widehat{\mathbf{F}}\} = \mathbb{B}\mathbb{V}_{\widehat{\boldsymbol{\theta}}_{[-1,-1]}}\mathbb{B}^T$. Taking $\widehat{\mathbf{F}} \pm 2\{\text{diag}(\text{var}\{\widehat{\mathbf{F}}\})\}^{1/2}$ gives approximate 95% empirical Bayesian confidence bands for $\mathbf{F}$.

These Bayesian intervals have a nice frequentist property "across the function": in repeated random experiments with the same F, the observed coverage probabilities averaged over the observation points will tend to be close to the nominal coverage level. This property was borne out in the simulation experiments of several articles including those by Wahba (1983) and Nychka (1988) for the case of smoothing splines and by Wood (2006a) for thin-plate regression splines. It will be examined for the FGAM through a simulation study in Section 4.2.

Depending on the application, a particular linear combination of the elements of $\widehat{\mathbf{F}}$ may be of interest. If we let $\mathbf{c}$ be a vector of the same length as $\widehat{\mathbf{F}}$, then we can also construct confidence bands of the form $\mathbf{c}^T\widehat{\mathbf{F}} \pm 2\{\mathbf{c}^T(\text{var}\{\widehat{\mathbf{F}}\})\mathbf{c}\}^{1/2}$. For example, this could be used to determine approximately whether two observed curves have significantly different effects on the response at a particular value of t. Under a null hypothesis of $H_0 : \boldsymbol{\theta} = \mathbf{0}$, $\widehat{\boldsymbol{\theta}}$ is unbiased and we can use the sandwich estimator for the variance, $\mathbb{V}_f = \mathbb{V}_{\widehat{\boldsymbol{\theta}}}\mathbb{Z}^T\widehat{\mathbb{W}}_M\mathbb{Z}\mathbb{V}_{\widehat{\boldsymbol{\theta}}}/\hat{\phi}$, to conduct approximate hypothesis tests for subsets of $\boldsymbol{\theta}$. For example, we can construct surfaces of approximate t-statistics by scaling the estimated surface values by the reciprocal of their standard error (the diagonal elements of $\mathbb{V}_f$).

For any pointwise transformation, $H_t(\cdot)$, of the predictor used (including $H_t(x) = x$), it is of interest to test whether $\partial^2/\partial h^2 F(h, t) = 0$ for all h and t, since this implies $F\{H_t(x), t\} = \beta(t)H_t(x)$ for some function $\beta(\cdot)$. Since derivatives of B-splines are simple to compute, an estimate of the second derivative of the surface and the Bayesian confidence intervals for the second derivative are easily obtained by replacing $\mathbb{B}_x$ in Equation (9) with evaluations of the second derivatives of the x-axis B-splines evaluated at the same points used for $\mathbb{B}_x$. While we cannot use our confidence bands for global inferences of this type, they do provide a rough heuristic for the desired test.

3.5 MULTIPLE PREDICTORS

Because of the modularity of penalized splines (Ruppert, Wand, and Carroll 2003), including multiple functional predictors as well as scalar predictors in the model is straightforward. Each additional functional predictor requires that two more smoothing parameters be selected. We will outline the procedure for the case of two functional covariates [say $X_1(\cdot)$, $X_2(\cdot)$] and one scalar covariate (say W). The model is $g\{E(Y_i|X_{i,1}, X_{i,2}, W_i)\} = \theta_0 + \int_{\mathcal{T}_1} F_1\{X_{i,1}(t), t\}dt + \int_{\mathcal{T}_2} F_2\{X_{i,2}(t), t\}dt + F_3(W_i)$, and both $X_1(\cdot)$ and $X_2(\cdot)$ can be transformed by their empirical cdfs. Further extensions are similar. As before, we use B-spline bases for both axes for both functional predictors and now also for W. One must also choose degrees of differencing to be used for each penalty. Let $\mathbb{Z}^{(1)}$ and $\mathbb{Z}^{(2)}$ denote the matrices of integrated tensor product B-splines for X_1 and X_2, respectively. Similarly, define $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ [see Equation (6)]. Let $\mathbb{B}^{(W)}$ be the matrix of W B-splines evaluated at the observed values of W and let $\boldsymbol{\theta}^{(W)}$ be the corresponding vector of B-spline coefficients for W. The penalty matrix for the smooth of W is given by $\mathbb{P}^{(W)} = \lambda_w \mathbb{D}_w^T \mathbb{D}_w$, where $\mathbb{D}_w$ is the differencing matrix for W and λ_w is its smoothing parameter. For identifiability, add the constraint $\mathbf{1}^T \mathbb{B}^{(W)} \boldsymbol{\theta}^{(W)} = 0$ (the usual constraint for each functional component in a standard additive model). We place the same constraint on both functional predictors as in the previous section. Thus, we have three total constraints. Construct

$$\mathbb{Z} = \left[\mathbf{1} \;\; \mathbb{B}^{(W)} \;\; \mathbb{Z}^{(1)} \;\; \mathbb{Z}^{(2)}\right], \quad \mathbb{P} = \text{diag}(0, \mathbb{P}^{(W)}, \mathbb{P}^{(1)}, \mathbb{P}^{(2)}), \text{ and } \boldsymbol{\theta} = (1, \boldsymbol{\theta}^{(W)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})^T.$$

To accommodate a linear effect of the covariate W, replace $\mathbb{B}^{(W)}$ in $\mathbb{Z}$ with the observed values of W and replace $\mathbb{P}^{(W)}$ with zero in the above formula for $\mathbb{P}$.

Note that it is also possible to have a linear effect for some functional predictors and additive effects for others; for example, a model of the form $g\{E(Y_i|X_i)\} = \theta_0 + f(W_i) + \int_{\mathcal{T}_1} \beta(t)X_{1i}(t)dt + \int_{\mathcal{T}_2} F\{X_{2i}(t), t\}dt$. Using the roughness penalty approach for estimating FLMs mentioned in Section 4.1, this can be implemented by making straightforward changes to $\mathbb{Z}^{(1)}$ and $\mathbb{P}^{(1)}$ (see Ramsay and Silverman 2005, chap. 15, for details).

4. SIMULATION EXPERIMENT

In this section, we perform simulations to assess the empirical performance of our FGAM. We first assess the ability of our FGAM to predict out-of-sample data in the Gaussian response case and compare its performance with several other functional regression models. Next, we examine the coverage properties of the empirical Bayesian confidence bands proposed in Section 3.4.

To generate the data, we created 1000 replicate datasets each consisting of N curves sampled at 200 equally spaced points in $[0, 1]$ as follows: Let $X_i(t) = \sum_{j=1}^{J} \gamma_j [Z_{1ij}\phi_{1j}(t) + Z_{2ij}\phi_{2j}(t)]$ where $Z_{hij} \sim N(0, 1)$, $\phi_{1j}(t) = \sqrt{2}\cos(\pi jt)$, $\phi_{1j}(t) = \sqrt{2}\sin(\pi jt)$, and $\gamma_j = \frac{2}{j}$; $h = 1, 2$; $i = 1, \ldots, N$; $j = 1, \ldots, J$. We consider two values for J, $J = 5$ and $J = 500$, the former resulting in much smoother predictor trajectories. We examine two cases for the true surface, $F(x, t)$, one where the FLM holds, $F(X(t), t) = \beta(t)X(t)$ and the other where it does not. For the linear true model, $F(x, t) = xt$. For the nonlinear true model, we use $F(x, t) = -0.5 + \exp[-(\frac{x}{5})^2 - (\frac{t-0.5}{0.3})^2]$, which looks like a hill or bivariate normal density.

The error variance changes with each sample so that the empirical signal to noise ratio (SNR) defined by $\text{SNR} = \frac{s_{\hat{y}}^2}{\sigma^2}$, where $s_{\hat{y}}^2 = \frac{1}{N-1}\sum_{i=1}^{N}[\int_{\mathcal{T}} F(X_i(t), t)\,dt - N^{-1}\sum_{i=1}^{N}\int_{\mathcal{T}} F(X_i(t), t)\,dt]^2$, remains constant. We consider the values $\text{SNR} = 1, 2, 4, 8$ in our simulations.

4.1 OUT-OF-SAMPLE PREDICTIVE PERFORMANCE

We fit FGAM and compare its out-of-sample predictive accuracy with three other popular functional regression models, the FLM, the kernel estimator by Ferraty and Vieu (2006), and the functional additive model (FAM) by Müller and Yao (2008). The coding used in our analyses was done in R (R Development Core Team 2011). The `fda` package (Ramsay et al. 2011) implements the standard tools of functional data analysis in R. As an initial step in fitting our model, the FLMs and the FAM, we use this package to smooth the data using B-spline basis functions and a roughness penalty with smoothing parameter chosen by GCV.

There are two main approaches for estimating the coefficient function $\beta(\cdot)$ for a FLM. The first uses smoothing or penalized splines and the second uses a functional principal component analysis (fPCA). We refer to these as FLM1 and FLM2, respectively.

These models can be fit in R using the `fda` package, more specifically, the functions `fRegress` for FLM1 and `pca.fd` for FLM2 (see Ramsay, Hooker, and Graves 2009, chap. 9, for computational details). For FLM1, we choose the smoothing parameter by minimizing GCV. For FLM2, we conduct a fPCA with a constant, light amount of smoothing and retain enough components for each simulation scenario to explain 90% of the total variability of the functional predictor. Once the scores are estimated, the final step to estimating FLM2 is fitting an unpenalized linear model in the scores.

To fit the FAM, we use the same number of principal component scores and the same estimation procedure as for FLM2. The difference comes in the next step, where a GAM is fit using the scores as predictors. To estimate the GAM, we use the default settings in the `mgcv` package and 11 basis functions for each additive term.

The final model we fit is described in detail by Ferraty and Vieu (2006, chap. 5). The response is predicted by the nonlinear operator $r(X) := E(Y|X)$. This operator is estimated by a functional extension of the Nadaraya-Watson kernel estimator:

$$\widehat{r}(X) = \frac{\sum_{i=1}^{N} Y_i\, K\left\{h^{-1}d(X, X_i)\right\}}{\sum_{i=1}^{N} K\left\{h^{-1}d(X, X_i)\right\}}, \tag{10}$$

where K is an asymmetrical kernel with bandwidth h and d is a semimetric. Continuity or Lipschitz continuity of the regression operator in the semimetric is assumed. We used the quadratic kernel, $K(u) = \frac{3}{4}(1-u^2)1_{[-1,1]}(u)$, and the semimetric $d(X_i, X_{i'}) = [\int_{\mathcal{T}}\{X_i(t) - X_{i'}(t)\}^2 dt]^{1/2}$. Code for fitting this model with automatic bandwidth selection can be obtained from: *http://www.math.univ-toulouse.fr/staph/npfda*. Note the differences in the assumptions and complexities of these three models: the simplest model assumes the response is linear in the functional predictor, the FGAM lessens the restrictions to additivity in the functional predictor, and the kernel estimator makes no restrictions on the form of the regression function other than continuity.

Each training set contained 67 curves and 33 curves were used for the test set. The performance of the models was measured by the out-of-sample RMSE $= [33^{-1}\sum_{i\in\{\text{test set}\}}(y_i - \widehat{y}_i)^2]^{1/2}$. We report results for both the FGAM fit to the original data and the FGAM fit after X has been transformed using the empirical cdf transformation given in Equation (5). In both cases, six cubic B-splines were used for the x-axis and seven cubic B-splines were used for the t-axis with second degree difference penalties for both axes. The tuning parameter, γ, for the GCV criterion (Equation (8)) was taken to be 1.0 in all cases. The `mgcv` package requires that the number of coefficients to estimate be less than the sample size, so we must have the product of the dimensions of the bases be less than the sample size minus one (for the intercept). The results of the simulations are summarized in Figure 2.

The figure reports the median RMSEs across the 1000 simulations for each scenario and model. We see that the FGAM loses little to the FLM in terms of predictive accuracy when the FLM is the true model and provides substantial improvements in the case when the FLM is not the true model. In fact, all the models perform quite similarly in the linear true model case with the exception of the Ferraty and Vieu model (Equation (10)) that performs considerably worse. In the nonlinear true model case, we see that fitting an FGAM to the transformed data performs slightly better than fitting an FGAM to the original curves and that in general the FGAM offers significant advantages over all the other models. As expected, the differences in performance between the different models become more pronounced as the fixed empirical SNR increases.

4.2 Bayesian Confidence Band Performance

We now assess the average coverage probabilities (ACP) of the confidence bands from Section 3.4. The observed ACP for the ith simulation is given by

$$\text{ACP} = \frac{1}{625}\sum_{j=1}^{25}\sum_{k=1}^{25} I\left\{F\left(x_j^{(i)}, t_k^{(i)}\right) \in C_{0.95}\left(x_j^{(i)}, t_k^{(i)}\right)\right\},$$

where $\{(x_j^{(i)}, t_k^{(i)});\ j,k = 1,\ldots,25\}$ are a subset of the $N \times 200$ observed $X(t)$ and t values for the ith simulation and $C_{0.95}(x_j^{(i)}, t_k^{(i)})\}$ is the entry of $\widehat{\mathbf{F}}^{(i)} \pm 2\{\text{diag}(\text{var}\{\widehat{\mathbf{F}}^{(i)}\})\}^{1/2}$ corresponding to $(x_j^{(i)}, t_k^{(i)})$. We consider two values for the sample size, $N = 100$ (combining the training and test sets from the previous section) and $N = 500$, the same true surfaces from the previous section, and two values for the empirical SNR, 2 and 4. For both the x and t axes, we use nine basis functions, cubic B-splines, and a second-order difference penalty.

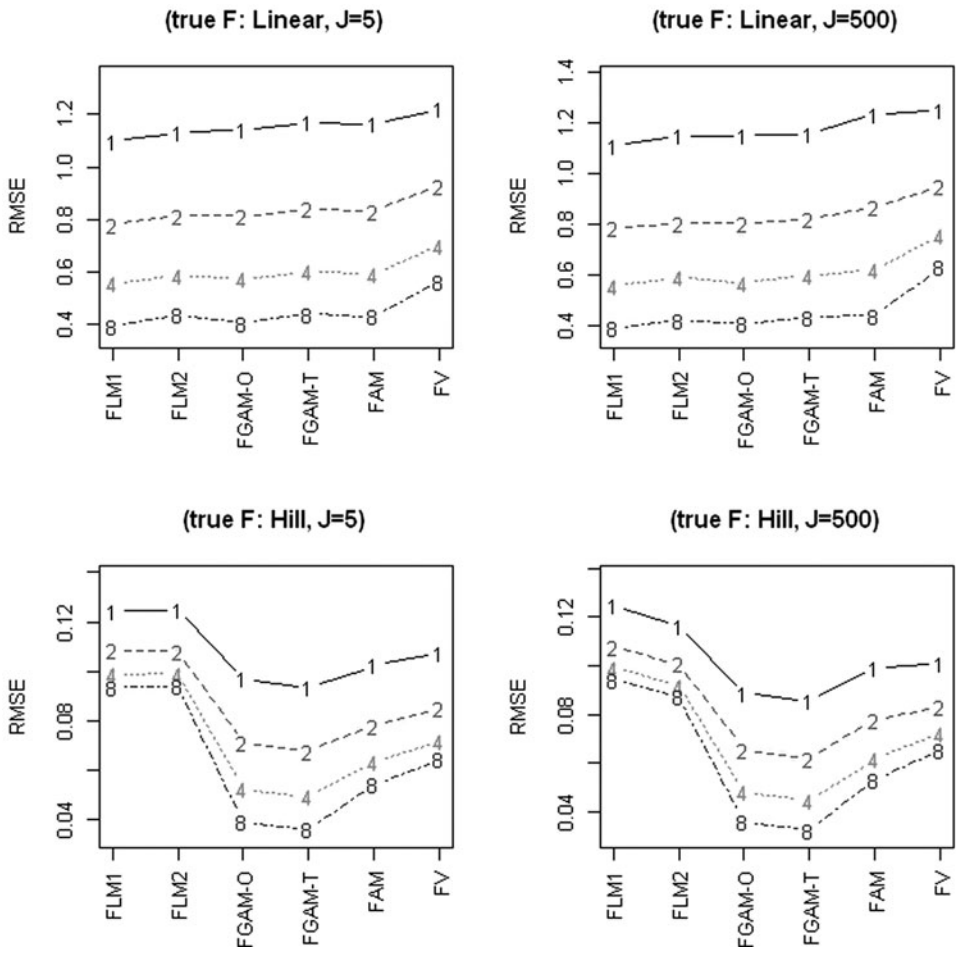

Figure 2. Median RMSE across 1000 simulations for six different functional regression models, four different empirical signal to noise ratios, and rough ($J = 500$) and smooth ($J = 5$) predictor functions. (a) Linear true model, (b) nonlinear ("Hill") true surface.

We report results for the FGAM fit without an intercept to the untransformed predictor curves with $J = 500$. The results for $J = 5$ were nearly identical.

To reduce the number of times that the confidence bands are evaluated at points outside the region jointly defined by the observed $(X_i(t_j), t_j)$ values, only grid points that are inside the convex hull defined by the observed values for each simulation are used in the calculation of mean ACP. A final modification is necessary to account for the identifiability constraint imposed on the FGAM. To do this, we fit the FGAM (including the constraint) with negligible amounts of smoothing to the true $E(Y_i|X_i)$ values (without noise) and take the fitted values to be the true responses. The mean ACP across the 500 simulations is displayed in Table 1 for each simulation scenario.

We see from the table that the coverage is fairly close to the nominal level of 0.95, though there is a slight problem with over-coverage in all the scenarios. Further analysis shows that the average estimated Bayesian standard errors for the surface are larger than the Monte

Table 1. Mean ACP across 500 simulations for nominal coverage probability 0.95

	$N = 100$		$N = 500$	
True surface	SNR = 2	SNR = 4	SNR = 2	SNR = 4
Linear	0.9746	0.9684	0.9704	0.9702
Nonlinear	0.9597	0.9665	0.9613	0.9592

Carlo standard deviation of the estimated surface, which is causing in the over-coverage. This is a biproduct of the Bayesian intervals trying to correct for the smoothing bias inherent in nonparametric regression. Recall that these intervals do not account for uncertainty in the estimation of λ_1 and λ_2. If more precise confidence bands are required, alternatives such as bootstrapping could be employed (see Wood 2006a, Sec. 4). These results indicate that it is safe to use the Bayesian confidence bands to make inferences about the true surface $F(x, t)$. We additionally ran a subset of these simulation scenarios while computing the confidence bands using the sandwich estimator of the variance of the estimated surface (results not included) and found there could be substantial under-coverage in the nonlinear true model case as a result of bias due to smoothing.

5. APPLICATION TO DIFFUSION TENSOR IMAGING DATASET

We now assess the performance of our model on a DTI tractography study. DTI is a technique for measuring the diffusion of water in tissue. Water diffuses differently in different types of tissue, and measuring these differences allows for detailed images to be obtained. Our dataset comes from a study comparing certain white matter tracts of multiple sclerosis (MS) patients with control subjects. MS is a central nervous system disorder that leads to lesions in the white matter of the brain that disrupts the ability of cells in the brain to communicate with each other. This dataset was previously analyzed by Goldsmith et al. (2011) and Greven et al. (2010).

The result of the DTI tractography is a 3×3 symmetric, positive definite matrix (equivalently, a three-dimensional ellipsoid) that describes diffusion at each desired location in the tract. We consider three functions of the estimated eigenvalues from these matrices: fractional anisotropy, parallel diffusivity, and perpendicular diffusivity. Fractional anisotropy measures the degree to which the diffusion is different in directions parallel and perpendicular to the tract, with zero indicating an isotrophic diffusion. More precisely, if the eigenvalues of the ellipsoid are given by $\lambda_1, \lambda_2, \lambda_3$, fractional anisotropy is equal to $[3\{(\lambda_1 - \bar{\lambda})^2 + (\lambda_2 - \bar{\lambda})^2 + (\lambda_3 - \bar{\lambda})^2\}/\{2(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)\}]^{1/2}$, where $\bar{\lambda} = (\lambda_1 + \lambda_2 + \lambda_3)/3$. Parallel (or axial or longitudinal) diffusivity is the largest eigenvalue of the ellipsoid. Perpendicular diffusivity is an average of the two smaller eigenvalues (see Mori 2007, for an overview of DTI).

Standard magnetic resonance imagining is used for diagnosing MS, but it is believed that the extra information provided by the tract profiles produced from DTI can be used to understand the disease process better. As an example of the types of effects we could investigate with our model, it has been found (Reich et al. 2007) that parallel diffusivity is

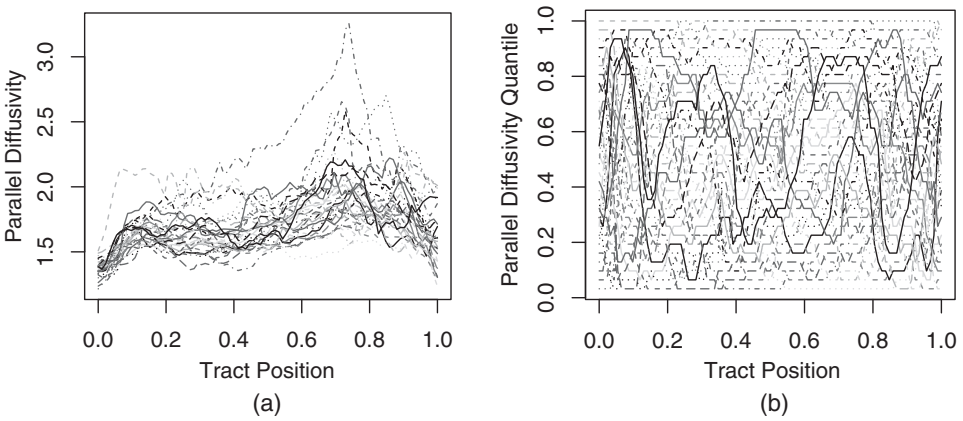

Figure 3. (a) Observed parallel diffusivity along the corpus callosum tract for a sample of MS patients. (b) Parallel diffusivity along the corpus callosum tract transformed by its empirical cdf for the same patients.

increased along the corticospinal tracts of people with MS. We would hope to see this effect if we were using parallel diffusivity measurements along that tract to predict MS status. We consider the corpus callosum tract in our analysis because it is related to cognition.

As an illustration of the FGAM, we fit our model using each of the three diffusion measures separately and compare the results with the same models introduced in the previous section. We also compare using the original curves as the predictor (Equation (1)) with using the empirical cdf of the curves (Equation (5)). Figure 3 contains plots of the parallel diffusivity measurements along the corpus callosum tract and the corresponding empirical cdf-transformed values for each subject in the training set.

Throughout the analysis, when fitting the FGAM, we use cubic B-splines with second-order difference penalties, six B-splines for the $x(p)$-axis, and seven B-splines for the t-axis. We found our results to be insensitive to these choices, and for brevity we do not include results for other values considered. Throughout this section, γ in (Equation (8)) is taken to equal 1.4. To evaluate the performance of the models, we examine their leave-one-curve-out prediction error. We repeatedly fit each model using all the samples except one and then use the fit to predict the left-out sample. This process is repeated until every sample has been left out once. Our performance measure is the root mean squared error, defined as $\text{RMSE} = [N^{-1}\sum_{i=1}^{N}(y_i - \widehat{y}_{(i)})^2]^{1/2}$, where $\widehat{y}_{(i)}$ is the predicted value of the ith response value when this sample is left out of the estimation.

5.1 PREDICTING PASAT SCORE

The first variable we predict is the result of a Paced Auditory Serial Addition Test (PASAT), a cognitive measure taking integer values between 0 and 60. The subject is given numbers at 3-sec intervals and asked to add the current number to the previous one. The final score is the total number of correct answers out of 60. MS patients often perform significantly worse than controls on this test. Since the corpus callosum is known to play a role in cognitive function, we might expect to see that the functional measurements along this tract have a significant impact in forecasting PASAT score. The PASAT was only

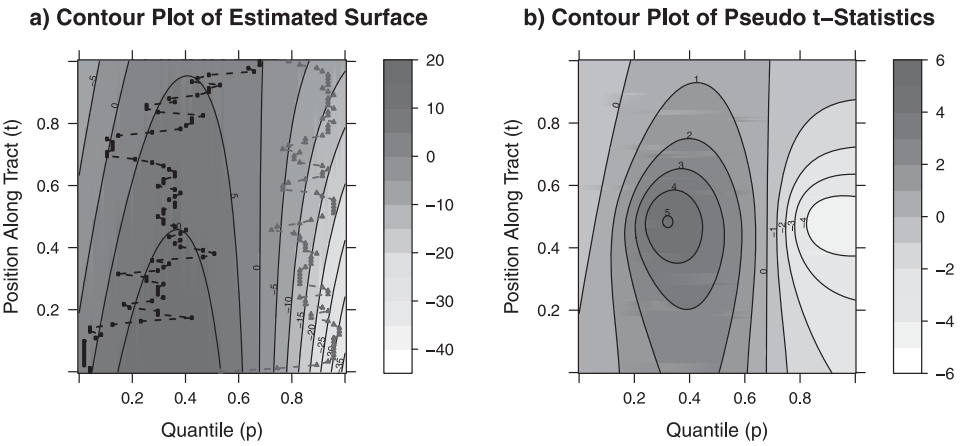

Figure 4. (a) Contour plot of the estimated surface, $\widehat{F}(p, t)$ [see (3)], for transformed parallel diffusivity along the corpus callosum tract. Also included are the transformed parallel diffusivity measurements for two subjects. (b) Contour plot of pseudo t-statistics (estimated surface value divided by its standard error). The response is PASAT score.

administered to subjects with MS. One subject with peculiar tract profiles was removed for simplicity and to avoid dealing with missing values.

The estimated surface $\widehat{F}(p, t)$ [see Equation (3)] is shown in Figure 4(a) for transformed parallel diffusivity. Figure 4(b) shows a contour plot of the observed pseudo-t statistics discussed in Section 3.4. We can see from this figure that parallel diffusivity for tract positions around 0.4–0.6 appears to be influential on the predicted response; subjects in the middle quantiles for this measurement at these positions are more likely to score higher on the PASAT, while the opposite is true for subjects in the upper quantiles at this location.

Figure 5 shows an example of a slice of the estimated surface when the untransformed curves are used for a fixed x value (left) and for a fixed position along the tract, t, (right). Parallel diffusivity along the corpus callosum is used as the predictor in these plots that also include twice standard error bands based on the sandwich estimator described earlier. Figure 5 also shows the same slices for the estimated second derivative of the surface with respect to t. This can give us a rough idea of whether the linear model is sufficient. In practice, we look at these plots for a representative sample of values with both the predictor value fixed and with the position fixed. We see that the second derivative is significantly nonzero in some regions, which suggests inadequacy of using an FLM in the untransformed predictors.

Table 2 reports out-of-sample RMSE from separately using each of the three different diffusivity measurements along the corpus callosum tract as predictors in the five models under consideration. Here, using FGAM with the empirical cdf transformation (FGAM-T) led to improved forecasting accuracy compared to using the raw measurements as predictors (FGAM-O). In fact, FGAM-T (Equation (5)) has lower out-of-sample RMSE than both FLMs for all the functional predictors considered, indicating that a linear model may be too restrictive in this application. Our FGAM-T compares favorably with the functional kernel regression model (Equation (10)) and the FAM, showing better performance when

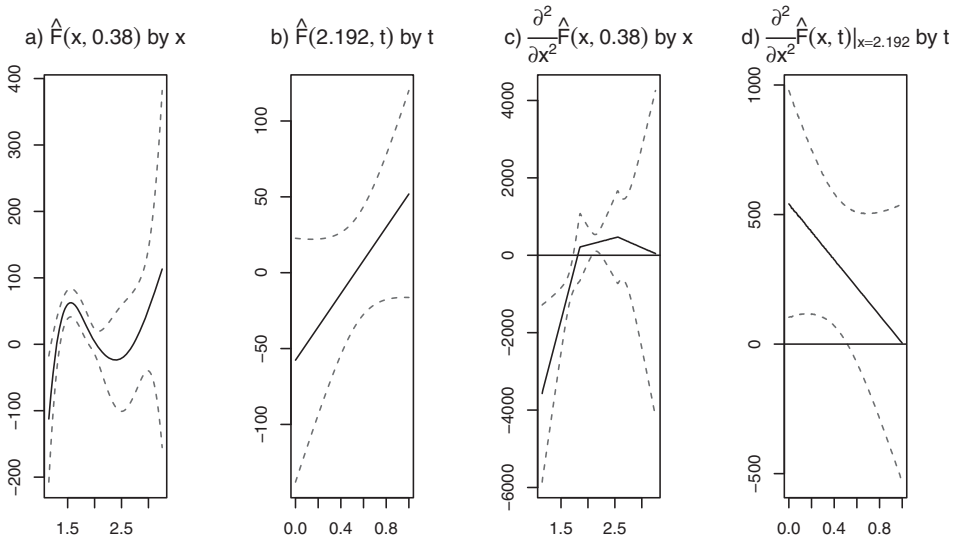

Figure 5. A sample of slices of the estimated surface [plots (a) and (b)] and estimated second derivative surface [(c) and (d)] for fixed tract positions [(a) and (c)] and fixed untransformed actual predictor [(b) and (d)] along with the corresponding Bayesian confidence bands for parallel diffusivity with PASAT score as the response variable.

either perpendicular diffusivity or fractional anisotropy are used as predictors. Though the kernel estimator provided slightly improved predictions in the parallel diffusivity case, the complex nature of its fit makes visualization difficult, so it is less useful than the FGAM for helping us understand the relationship between the DTI measurements and the PASAT scores.

5.2 PREDICTING MS STATUS: LOGISTIC LINK

We now consider classifying the disease status of subjects. Since the PASAT was only given to the subjects with MS, our sample size is now 88 and includes controls. We include results using the untransformed curves only. The results using the quantile transformation were similar. We again use the leave-one-curve-out procedure described earlier. Fitting the FGAM resulted in the estimated surface displayed in Figure 1 when perpendicular diffusivity is used as the predictor. The observed perpendicular diffusivity for two subjects is overlaid on the plot; recall the interpretation given in the Introduction. It appears that

Table 2. Leave-one-curve-out RMSEs for the three different functional predictors of PASAT score using the following models: FGAM using the original curves (FGAM-O), FGAM using the empirical cdf transformation [FGAM-T, (5)], FLM1, FLM2, FV (10), and FAM

Measurement	FGAM-O	FGAM-T	FLM1	FLM2	FV	FAM
Perpendicular diffusivity	12.22	10.46	10.98	11.27	11.16	11.71
Fractional anisotropy	12.55	11.60	11.87	11.91	12.11	12.70
Parallel diffusivity	11.94	12.09	12.32	12.24	11.97	11.86

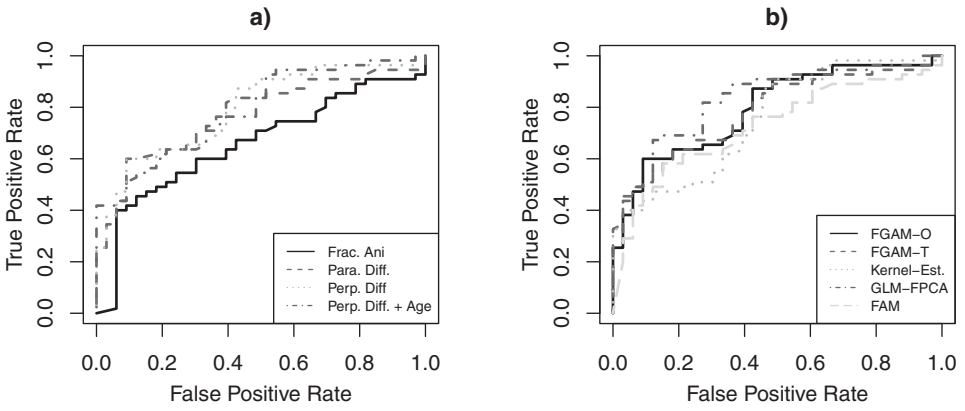

Figure 6. (a) Leave-one-curve-out ROC curves for different FGAMs fit each using a different functional predictor and an FGAM including perpendicular diffusivity and a functional component for age. The response is MS status. (b) Leave-one-curve-out ROC curves for both FGAM fits, and three other functional regression models when perpendicular diffusivity is the functional predictor.

the predictor values at the end of the tract corresponding to $t = 1$ have a strong influence in predicting disease status. Subjects in the lower range for perpendicular diffusivity at this end of the tract seem to be less likely to be classified as having MS, whereas subjects in the upper range at this position are more likely to have MS. Models were also fit using fractional anisotropy and parallel diffusivity as predictors. A fourth model was considered that included a nonparametric component for the subject's age in addition to using perpendicular diffusivity. Figure 6 contains a plot of the ROC curves for these fitted models. The model using fractional anisotropy performs almost universally worse than the other three models. None of the other three models considered perform universally better than the others. Including age as a covariate in the model with perpendicular diffusivity did not improve performance.

We also compared the FGAM fits to three other generalized functional regression models. The first is the Ferraty + Vieu estimator (Equation (10)) from the previous section. The use of this estimator for classification is discussed in detail by Ferraty and Vieu (2006, chap. 8). The second alternative model considered is a GLM in the functional principal component scores (GLM-FPCA) and the third model is a GAM in the functional principal component scores (FAM). The leave-one-curve-out ROC curves are displayed in the right plot of Figure 6 when perpendicular diffusivity is the covariate. There is little difference in performance between the models used.

6. DISCUSSION

A new model for functional regression with a scalar response has been developed. The functional linear model has been extended to an additive structure allowing for more complicated relationships to be modeled while still being highly interpretable. Our approach can handle responses from any exponential family distribution as well as multiple functional or scalar predictors.

In our simulation results, we showed that our FGAM can provide nearly identical results to the FLM when the FLM is the true model, and offered substantial improvements when the FLM was not the true model. We also showed that our proposed confidence bands can achieve average coverage probabilities close to the nominal confidence level. For the analysis of the DTI dataset, FGAM performed favorably when compared with some standard functional regression models.

Our methodology opens up many research problems. One goal for future work is to add several more functional predictors (e.g., for the DTI dataset, multiple tracts and more summaries of the diffusion for each tract). This would require faster techniques for smoothing parameter selection than our current methods. A Bayesian model and MCMC could have also been used. Alternatively, the recent work by Wood (2011) advocates the use of generalized linear mixed models (GLMMs) estimated by restricted maximum likelihood for smoothing parameter selection. Our analysis of the DTI dataset did not consider the longitudinal nature of the study. The use of a GLMM to incorporate random effects would allow us to model these extra subject visits in a manner similar to Goldsmith et al. (2010). Datasets with this type of structure are becoming more and more common (e.g., see the analysis by Di et al. 2009, of the Sleep Healthy Heart Study).

There are multiple alternatives to the Bayesian approach we used for obtaining approximate confidence bands for the estimated surface $\widehat{\mathbf{F}}$. Ruppert, Wand, and Carroll (2003, chap. 6) provided an overview of some of them. Bootstrap procedures are commonly used and have been developed for functional nonparametric regression (Ferraty, Van Keilegom, and Vieu 2010). Fahrmeir and Lang (2001) used a Bayesian approach involving Markov random field priors with estimation performed using MCMC. We note that with the approach we use, it is also possible to obtain confidence intervals for nonlinear functions of the model parameters (see Wood 2006a).

Also of interest for further work would be obtaining formal tests of the additivity and linearity assumptions for our model and the FLM, respectively, and convergence rates for $\widehat{\mathbf{F}}$. Recent results for P-splines suggest we can obtain theoretical results for the Riemann sum approximation to our model. Li (2011) showed the equivalence of a P-spline estimator to a backfitting projection algorithm and used this result to obtain asymptotic results for the P-spline estimator for the case of piecewise constant or linear splines with first- or second-order difference penalties. Care must be taken to deal with the high degree of multicollinearity among the covariates; the assumptions placed on the probabilistic structure of the functional predictors will likely have an important role in obtaining rates of convergence.

SUPPLEMENTARY MATERIALS

Code for fitting the FGAM in R is available in the package `refund`. Additional code for conducting the simulations and with additional examples is available in `fgam.zip`. Also included is `FGAMappendix.pdf` that contains further details on identifiability and a description of how to fit the FGAM using the `mgcv` package. See the README file for a detailed description of each file.

ACKNOWLEDGMENTS

This work was completed while Mathew W. McLean was a Ph.D. student in the School of Operations Research and Information Engineering at Cornell University. He was supported by an NSERC PGS-D award and by award number R01NS060910 from the National Institute of Neurological Disorders and Stroke. Giles Hooker was partially supported from NSF grants DEB-0813743, CMG-0934735, and DMS-1053252. David Ruppert was partially supported by award number R01NS060910 from the National Institute of Neurological Disorders and Stroke and grant DMS-0805975 from the National Science Foundation. Fabian Scheipl was partially supported by the German Research Foundation through the Emmy Noether Programme, grant GR 3793/1-1 to Sonja Greven. Staicu's research was supported by U.S. National Science Foundation grant number DMS 1007466. We thank Ciprian Crainiceanu, Daniel Reich, the National Multiple Sclerosis Society, and Peter Calabresi for the DTI dataset, and Joyjit Roy and Martin Larsson for useful discussions. We also thank two referees and the associate editor for helpful comments.

[Received May 2011. Revised August 2012.]

REFERENCES

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," *The Annals of Statistics*, 17, 453–510. [250]

Cardot, H., Ferraty, F., and Sarda, P. (2003), "Spline Estimators for the Functional Linear Model," *Statistica Sinica*, 13, 571–592. [253]

Crambes, C., Kneip, A., and Sarda, P. (2009), "Smoothing Splines Estimators for Functional Linear Regression," *The Annals of Statistics*, 37, 35–72. [253]

Di, C. Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009), "Multilevel Functional Principal Component Analysis," *Annals of Applied Statistics*, 3, 458–488. [267]

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing With B-Splines and Penalties," *Statistical Science*, 11, 89–121. [251,253]

Fahrmeir, L., and Lang, S. (2001), "Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors," *Journal of the Royal Statistical Society,* Series C, 50, 201–220. [267]

Ferraty, F., Van Keilegom, I., and Vieu, P. (2010), "On the Validity of the Bootstrap in Non-Parametric Functional Regression," *Scandinavian Journal of Statistics*, 37, 286–306. [267]

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, New York: Springer Verlag. [259,266]

Goldsmith, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2010), *Longitudinal Penalized Functional Regression*, Technical Report, Baltimore, MD: Department of Biostatistics, Johns Hopkins University. [267]

Goldsmith, J., Feder, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011), "Penalized Functional Regression," *Journal of Computational and Graphical Statistics*, 20, 830–851. [253,262]

Greven, S., Crainiceanu, C. M., Caffo, B., and Reich, D. (2010), "Longitudinal Functional Principal Component Analysis," *Electronic Journal of Statistics*, 4, 1022–1054. [262]

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, Boca Raton, FL: Chapman & Hall/CRC. [250,257]

James, G. (2002), "Generalized Linear Models With Functional Predictors," *Journal of the Royal Statistical Society,* Series B, 64, 411–432. [250]

James, G., and Silverman, B. W. (2005), "Functional Adaptive Model Estimation," *Journal of the American Statistical Association*, 100, 565–577. [250]

Li, E., Wang, N., and Wang, N. Y. (2007), "Joint Models for a Primary Endpoint and Multiple Longitudinal Covariate Processes," *Biometrics*, 63, 1068–1078. [251]

Li, Y. (2011), "Aspects of Penalized Splines," Ph.D. thesis, Cornell University. [267]

Mammen, E., Linton, O., and Nielsen, J. (1999), "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm Under Weak Conditions," *The Annals of Statistics*, 27, 1443–1490. [250]

Marx, B. D., and Eilers, P. H. C. (1998), "Direct Generalized Additive Modeling With Penalized Likelihood," *Computational Statistics and Data Analysis*, 28, 193–209. [251,253,254,257]

——— (2005), "Multidimensional Penalized Signal Regression," *Technometrics*, 47, 13–22. [255]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC. [256]

Mori, S. (2007), *Introduction to Diffusion Tensor Imaging*, Oxford: Elsevier Science. [262]

Müller, H. G., and Stadtmüller, U. (2005), "Generalized Functional Linear Models," *The Annals of Statistics*, 33, 774–805. [250]

Müller, H. G., and Yao, F. (2008), "Functional Additive Models," *Journal of the American Statistical Association*, 103, 1534–1544. [249,250,251,259]

Nielsen, J. P., and Sperlich, S. (2005), "Smooth Backfitting in Practice," *Journal of the Royal Statistical Society,* Series B, 67, 43–61. [251]

Nychka, D. (1988), "Bayesian Confidence Intervals for Smoothing Splines," *Journal of the American Statistical Association*, 83, 1134–1143. [257]

Opsomer, J. D., and Ruppert, D. (1997), "Fitting a Bivariate Additive Model by Local Polynomial Regression," *The Annals of Statistics*, 25, 186–211. [250]

——— (1998), "A Fully Automated Bandwidth Selection Method for Fitting Additive Models," *Journal of the American Statistical Association*, 93, 605–619. [250]

——— (1999), "A Root-n Consistent Backfitting Estimator for Semiparametric Additive Modeling," *Journal of Computational and Graphical Statistics*, 8, 715–732. [250]

R Development Core Team. (2011), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing. Available at *http://www.R-project.org/*. [259]

Ramsay, J. O., and Dalzell, C. J. (1991), "Some Tools for Functional Data Analysis," *Journal of the Royal Statistical Society,* Series B, 53, 539–572. [250]

Ramsay, J. O., Hooker, G., and Graves, S. (2009), *Functional Data Analysis With R and MATLAB*, New York: Springer. [251,259]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer. [258]

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2011), *fda: Functional Data Analysis*, available at *http://CRAN.R-project.org/package=fda*. R package version 2.2.6. [259]

Reich, D. S., Smith, S. A., Zackowski, K. M., Gordon-Lipkin, E. M., Jones, C. K., Farrell, J. A. D., Mori, S., van Zijl, P., and Calabresi, P. A. (2007), "Multiparametric Magnetic Resonance Imaging Analysis of the Corticospinal Tract in Multiple Sclerosis," *Neuroimage*, 38, 271–279. [262]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, New York: Cambridge University Press. [251,258,267]

Wahba, G. (1983), "Bayesian 'Confidence Intervals' for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society,* Series B, 45, 133–150. [257]

Wood, S. N. (2004), "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models," *Journal of the American Statistical Association*, 99, 673–686. [256]

——— (2006a), "On Confidence Intervals for Generalized Additive Models Based on Penalized Regression Splines," *Australian and New Zealand Journal of Statistics*, 48, 445–464. [257,262,267]

——— (2006b), *Generalized Additive Models: An Introduction with R*, Boca Raton, FL: CRC Press. [251,253,256,257]

——— (2011), "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semi-parametric Generalized Linear Models," *Journal of the Royal Statistical Society,* Series B, 73, 3–36. [256,267]

Yuan, M., and Cai, T. T. (2010), "A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression," *The Annals of Statistics*, 38, 3412–3444. [253]