Journal of the
Royal Statistical Society

SERIES C
Applied
Statistics

Appl. Statist. (2012)
61, *Part* 3, *pp.* 453–469

Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements

Jeff Goldsmith,

Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

Ciprian M. Crainiceanu,

Johns Hopkins University, Baltimore, USA

Brian Caffo

Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

and Daniel Reich

National Institutes of Health, Bethesda, USA

[Received January 2011. Final revision September 2011]

Summary. We describe and analyse a longitudinal diffusion tensor imaging study relating changes in the microstructure of intracranial white matter tracts to cognitive disability in multiple-sclerosis patients. In this application the scalar outcome and the functional exposure are measured longitudinally. This data structure is new and raises challenges that cannot be addressed with current methods and software. To analyse the data, we introduce a penalized functional regression model and inferential tools designed specifically for these emerging types of data. Our proposed model extends the generalized linear mixed model by adding functional predictors; this method is computationally feasible and is applicable when the functional predictors are measured densely, sparsely or with error. On-line supplements compare two implementations, one likelihood based and the other Bayesian, and provide the software that is used in simulations; the likelihood-based implementation is included as the lpfr() function in the R package refund that is available in the Comprehensive R Archive Network.

Keywords: Bayesian inference; Functional regression; Mixed models; Smoothing splines

1. Introduction

Traditionally, longitudinal studies have collected scalar measurements on subjects over time. As technologies for the collection and storage of larger measurements have become widely available, longitudinal studies have begun to collect functional or imaging observations on subjects over several visits. One example is our current data set, in which diffusion tensor imaging (DTI) brain scans are recorded for many multiple-sclerosis (MS) patients over several visits with the goal of assessing the effect of neurodegeneration on disability. Adequately relating functional predictors to accompanying scalar outcomes requires longitudinal functional regression models,

Address for correspondence: Jeff Goldsmith, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21210, USA.
E-mail: jgoldsmi@jhsph.edu

0035–9254/12/61453

which are not currently available. We address this problem by introducing a generally applicable regression model that adds subject-specific random effects to the well-studied cross-sectional functional regression model. We also develop inferential techniques for all parameters in this new model and implement these methods in computationally efficient and publicly available software that is available in the `refund` R package.

1.1. Description of the data

Our application explores the relationship between cerebral white matter tracts in MS patients and cognitive impairment over time. White matter tracts are made up of myelinated axons: axons are the long projections of a neuron that transmit electrical signals and myelin is a fatty substance that surrounds the axons in white matter, enabling the electrical signals to be carried very quickly. MS, which is an autoimmune disease which results in axon demyelination and lesions in white matter tracts, leads to significant disability in patients.

DTI is a magnetic resonance imaging based modality that traces the diffusion of water in the brain. Because water diffuses anisotropically in the white matter and isotropically elsewhere, DTI is used to generate images of the white matter specifically (Basser *et al.*, 1994, 2000; LeBihan *et al.*, 2001; Mori and Barker, 1999). Several measurements of water diffusion are provided by DTI, including fractional anisotropy and mean diffusivity. Continuous summaries of white matter tracts, parameterized by distance along the tract and called tract profiles, can be derived from diffusion tensor images.

By collecting longitudinal information about patient cognitive function and about disease progression via DTI, researchers hope to understand the relationship between MS and disability better. From this study, we have densely sampled mean and parallel diffusivity measurements from several white matter tracts. Our data set consists of 100 subjects, 66 women and 34 men, aged between 21 and 70 years at first visit. The number of visits per subject ranged from 2 to 8, with a median of 3, and were approximately annual; a total of 340 visits were recorded. At each visit full DTI scans were obtained and used to create tract profiles, accompanied by several tests of cognitive and motor function with scalar outcomes.

In Fig. 1 we display a functional predictor and cognitive disability outcome for two subjects over time. We stress that this data structure, with high dimensional predictors and scalar outcomes observed longitudinally, is increasingly common. Moreover, we emphasize that our methods are motivated by this study but are generally applicable. A single-level analysis of these data was presented in Goldsmith *et al.* (2011b).

1.2. Model proposed

More formally, we consider the setting in which we observe for each subject $1 \leqslant i \leqslant I$ at each visit $1 \leqslant j \leqslant J_i$ data of the form $[Y_{ij}, W_{ij1}(s), \ldots, W_{ijK}(s), X_{ij}]$, where Y_{ij} is a scalar outcome, $W_{ijk}(s) \in \mathcal{L}^2[0,1]$, $1 \leqslant k \leqslant K$, are functional covariates and X_{ij} is a row vector of scalar covariates. We propose the longitudinal functional regression outcome model

$$\begin{aligned} Y_{ij} &\sim \mathrm{EF}(\mu_{ij}, \eta), \\ g(\mu_{ij}) &= X_{ij}\beta + Z_{ij}b_i + \sum_{k=1}^{K} \int_0^1 W_{ijk}(s)\,\gamma_k(s)\,\mathrm{d}s \end{aligned} \tag{1}$$

where '$\mathrm{EF}(\mu_{ij}, \eta)$' denotes an exponential family distribution with mean μ_{ij} and dispersion parameter η. Here $X_{ij}\beta$ is the standard fixed effects component, $Z_{ij}b_i$ is the standard random-effects component, $b_i \sim N(0, \sigma_{\mathbf{b}}^2 \mathbf{I}_I)$ are subject-specific random effects and the $\int_0^1 W_{ijk}(s)\,\gamma_k(s)\,\mathrm{d}s$ are the functional effects. Both the functional coefficients $\gamma_k(s)$ and the scalar coefficients β are population level parameters, rather than subject-specific effects, and do not vary across visits.

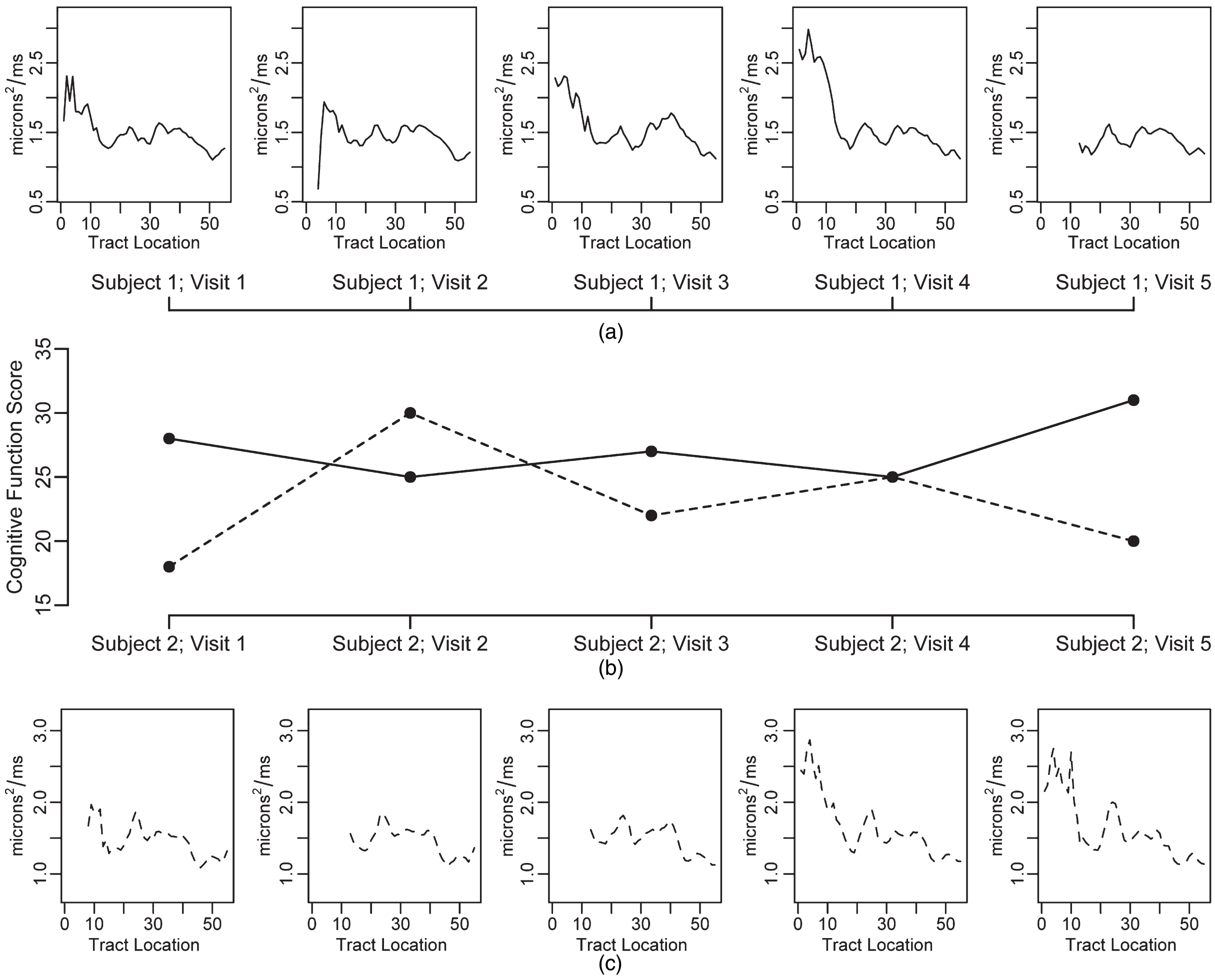

Fig. 1. Structure of the data: (a), (c) parallel diffusivity tract profiles of the right corticospinal tract corresponding to each subject–visit outcome, which we use as a functional regressor; (b) scalar paced auditory serial addition test cognitive disability measure for two subjects over five visits (measured as the number of questions answered correctly out of 60)

Model (1) is novel in that it adds subject-specific random effects to the standard cross-sectional functional regression model; from another point of view, this model adds functional predictors to generalized linear mixed models. The latter viewpoint is particularly instructive in that ideas which are familiar from traditional longitudinal data analysis can be transferred seamlessly to longitudinal functional data analysis. As an example, many different structures of random effects b_i, including random intercepts and slopes, will be needed in practice and can be implemented by appropriately specifying the random-effect design matrices and covariance structures. For expositional clarity we use a single vector of subject-specific random effects and assume that these random effects are independent, but more complex structures can be included. In practice the $W_{ijk}(s)$ are not truly functional but are observed on a dense (or sparse) grid and often with error; moreover, the predictors are not necessarily observed over the same domain. In fact, whereas the measurement error is negligible in our application, there are some missing values in the functional predictor in Fig. 1, and the mean and parallel diffusivity functional predictors have different domains. We point out that solutions to these problems are well known (Di *et al.*, 2009; Ramsay and Silverman, 2005; Staniswalis and Lee, 1998; Yao *et al.*, 2003) and we do not discuss them here.

Our proposed approach is based on the single-level functional regression model that is developed in Goldsmith *et al.* (2011a) but is extended to the scientifically important longitudinal setting. Strengths of this approach are that

(a) it extends functional regression to model the association between outcomes and functional predictors when observations are clusted into groups or subjects,
(b) it casts a novel functional model in terms of well-understood mixed models,
(c) it is applicable in any situation in which the $W_{ijk}(s)$ are observed or can be estimated, including when they are observed sparsely or with error,
(d) it can be fitted by using standard, well-developed statistical software available as the `lpfr()` function in the `refund` R package,
(e) from a Bayesian perspective, it allows the joint modelling of longitudinal outcomes and predictors, and
(f) it provides confidence or credible intervals for all the parameters of the model, including the functional parameters.

We emphasize this final point, as confidence intervals are rarely discussed in the functional regression literature, and in penalized approaches to functional regression they are typically bootstrap or empirical intervals. The connection to mixed models provides a simple, statistically principled approach for constructing confidence or credible intervals.

1.3. Existing methods for functional regression

We contrast the setting of this paper with the large body of existing functional regression work. Foremost, existing functional regression work (Cardot *et al.*, 1999, 2003, 2007; Cardot and Sarda, 2005; Crambes *et al.*, 2009; Ferraty and Vieu, 2006; James, 2002; James *et al.*, 2009; Marx and Eilers, 1999; Müller and Stadtmüller, 2005; Ramsay and Silverman, 2005; Reiss and Ogden, 2007) deals only with cross-sectional regression. Here, we are focused on longitudinally observed functional predictors and outcomes, which necessitate the addition of random effects to the standard functional regression model. To the best of our knowledge, this setting has not been considered previously. Moreover, the addition of random effects increases the complexity of the functional regression model, requiring new methodology implemented in efficient software.

Additionally, we point out that several alternative penalized approaches to single-level functional regression exist (Cardot *et al.*, 2003; Cardot and Sarda, 2005; Reiss and Ogden, 2007), but that each of these incorporates a computationally expensive cross-validation procedure. How and whether such methods would extend to fitting longitudinal models of type (1) remains to be elucidated. The additional complexity of a longitudinal model may increase the computational burden, perhaps prohibitively. Low dimension approaches to single-level functional regression (Müller and Stadtmüller, 2005), in which the smoothness of $\gamma_k(s)$ depends on the dimension of its basis, are less automated than penalized approaches but are, notably, generalizable to the longitudinal setting. In fact, such models are a special case of the method that we propose in Section 2.

It is also important to distinguish the proposed longitudinal functional regression model from the well-developed functional analysis-of-variance models (Brumback and Rice, 1998; Guo, 2002). Here, one observes functions organized into groups; the goal is to express the functions as a combination of a group mean, a subject-specific deviation from the group mean and possibly a subject-visit-specific deviation. Much of this work has focused on the estimation of and inference for the group means and has employed penalized splines in expressing these functions. A scalar-on-function regression that incorporates random effects to account for group effects, as proposed in this paper, is not considered; we, however, point out that the work in functional analysis of variance suggests that numerous data sets necessitating such a model exist and are under investigation. Others (Di *et al.*, 2009; Greven *et al.*, 2010) have extended functional principal components analysis to the multilevel and longitudinal settings, emphasizing parsimonious and computationally efficient methods for the expression of subject-visit-specific curves. Although these methods are the state of the art in describing the variability in observed multilevel and longitudinal functions, they do not consider accompanying longitudinal outcomes.

Finally, we highlight the distinction between the treatment of scalars observed longitudinally as sparse functional covariates (Hall *et al.*, 2006; Müller, 2005; Yao *et al.*, 2005) and the current setting, in which functions are observed longitudinally.

The introduction of a longitudinal functional regression model, therefore, fills a gap in the functional data analysis literature. Whereas the approach proposed is based on the framework that is described in Goldsmith *et al.* (2011a) and arises naturally therein, the longitudinal extension that is developed here defines a broadly useful class of functional regression models. Moreover, computationally feasible software for the estimation and inference related to longitudinal functional regression models is freely available on line.

The remainder of this paper is organized in the following way. Section 2 describes the proposed general method for longitudinal functional regression. In Section 3 we pursue a simulation study to examine the viability of the method proposed and in Section 4 we apply our method to the DTI data. We end with a discussion in Section 5. Implementations of the method that is proposed in Section 2, provided in both likelihood-based and Bayesian frameworks and accompanied by a discussion of the advantages and disadvantages of each, are available in an on-line appendix and all the software that is used in the simulation exercise is available from

`http://www.blackwellpublishing.com/rss`

2. Longitudinal penalized functional regression

The longitudinal penalized functional regression method builds on an approach to single-level functional regression in which the penalization is achieved via a mixed model (Goldsmith *et al.*, 2011a). Briefly, the two steps in this method are

(a) to express the predictors by using a large number of functional principal components obtained from a smooth estimator of the covariance matrix estimator and
(b) to express the function coefficient by using a penalized spline basis.

The addition of random effects to this model arises naturally and with a minimal increase in complexity. In what follows, we shall use i to index subject, j to index visit, k to index functional predictor, l to index objects associated with principal component bases and m to index objects that are associated with spline bases.

Specifically, given data of the form $[Y_{ij}, W_{ij1}(s), \ldots, W_{ijK}(s), X_{ij}]$ for subjects $1 \leqslant i \leqslant I$ over visits $1 \leqslant j \leqslant J_i$, we model the functional effect in the following way. First, express the $W_{ijk}(s)$ in terms of a truncated Karhunen–Loève decomposition, i.e. let $\Sigma^{\mathbf{W}_k}(s,t) = \text{cov}\{W_{ijk}(s), W_{ijk}(t)\}$ be the covariance operator on the kth observed functional predictor; thus $\Sigma^{\mathbf{W}_k}(s,t)$ is a bivariate function providing the covariance between two locations of the functional predictor. Further, let $\Sigma_{l=1}^{\infty} \lambda_{kl}\, \psi_{kl}(s)\, \psi_{kl}(t)$ be the spectral decomposition of $\Sigma^{\mathbf{W}_k}(s,t)$, where $\lambda_{k1} \geqslant \lambda_{k2} \geqslant \ldots$ are the non-increasing eigenvalues and $\boldsymbol{\psi}_k(\cdot) = \{\psi_{kl}(\cdot) : l \in \mathbb{Z}^+\}$ are the corresponding orthonormal eigenfunctions. In practice, functional predictors are observed over finite grids, and often with error. Thus we estimate $\Sigma^{\mathbf{W}_k}(s,t)$ by using a method-of-moments approach, and then smooth the off-diagonal elements of this observed covariance matrix to remove the 'nugget effect' that is caused by measurement error (Staniswalis and Lee, 1998; Yao *et al.*, 2003). Moreover, in the case that the $W_{ijk}(s)$ are sampled sparsely or over different grids, we can construct an estimate of the covariance matrix by using the following two-stage procedure (Di *et al.*, 2009). First, use a fine grid to bin each subject's observations to construct a rough estimate of the covariance matrix based on these undersmoothed functions; second, smooth the rough covariance matrix that is estimated in the previous step.

A truncated Karhunen–Loève approximation for $W_{ijk}(s)$ is given by

$$W_{ijk}(s) = \mu_k(s) + \sum_{l=1}^{K_w} c_{ijkl}\, \psi_{kl}(s),$$

where K_w is the truncation lag, the $c_{ijkl} = \int_0^1 \{W_{ijk}(s) - \mu_k(s)\}\, \psi_{kl}(s)\, \mathrm{d}s$ are uncorrelated random variables with variance λ_{kl} and $\mu_k(s)$ is the mean of the kth functional predictor, taken over all subjects and visits. Unbiased estimators of c_{ijkl} can be obtained either as the Riemann sum approximation to the integral $\int_0^1 \{W_{ijk}(s) - \mu_k(s)\}\, \psi_{kl}(s)\, \mathrm{d}s$ or via the mixed effects model (Crainiceanu *et al.*, 2009; Di *et al.*, 2009):

$$\left.\begin{aligned} W_{ijk}(s) &= \mu_k(s) + \sum_{k=1}^{K_w} c_{ijkl}\, \psi_{kl}(s) + \varepsilon_{ijk}(s), \\ \mathbf{c}_{ijk} &\sim N(0, \Lambda_k), \\ \varepsilon_{ijk}(s) &\sim N(0, \sigma^2_{\mathbf{W}_k}), \end{aligned}\right\} \tag{2}$$

where $\mathbf{c}'_{ijk} = (c_{ijk1}, \ldots, c_{ijkK_w})$ is the vector of subject–visit-specific loadings for the kth functional predictor, Λ_k is a $K_w \times K_w$ matrix with (l,l)th entry λ_{kl} and 0 elsewhere, and the $\mathbf{c}_{ijk}$ and $\varepsilon_{ijk}(s)$ are mutually independent for every i, j and k. Note that we have expressed the $W_{ijk}(s)$ without taking the repeated subject level observations into account; alternatively, we could use longitudinal functional principal components analysis for a decomposition that borrows strength across visits (Greven *et al.*, 2010).

The second step in longitudinal penalized functional regression is to model the coefficient functions $\gamma_k(s)$ by using a large spline basis with smoothness induced explicitly via a mixed effects model. For example, let $\boldsymbol{\phi}_k(s) = \{\phi_{k1}(s), \phi_{k2}(s), \ldots, \phi_{kK_g}(s)\}$ be a truncated power series spline basis, so that

$$\gamma_k(s) = \boldsymbol{\phi}_k(s)\boldsymbol{g}_k = g_{k1} + g_{k2}t + \sum_{m=3}^{K_g} g_{km}(t - \kappa_{km})_+$$

where $\boldsymbol{g}_k=(g_{k1},\dots,g_{kK_g})^{\mathrm{T}}$ and $\{\kappa_{km}\}_{m=3}^{K_g}$ are knots (note that the index on the knots begins at 3 to match the index of the spline terms in $\boldsymbol{g}_k$). Thus,

$$\int_0^1 W_{ijk}(s)\,\gamma_k(s)\,\mathrm{d}s=a_k+\int_0^1 \mathbf{c}'_{ijk}\,\boldsymbol{\psi}_k^{\mathrm{T}}(s)\,\boldsymbol{\phi}_k(s)\boldsymbol{g}_k\,\mathrm{d}s=a_k+\mathbf{c}'_{ijk}\mathbf{M}_k\boldsymbol{g}_k,$$

where $\mathbf{M}_k$ is a $(K_w\times K_g)$-dimensional matrix with the (l,m)th entry equal to $\int_0^1 \psi_{kl}(s)\,\phi_{km}(s)\,\mathrm{d}s$ and $a_k=\int_0^1 \mu_k(s)\,\gamma_k(s)\,\mathrm{d}s$. Then, letting $\mathbf{C}_k$ be the $(\Sigma_{i=1}^I J_i)\times K_w$ matrix of principal component loadings with rows $\mathbf{c}'_{ijk}$, X be the design matrix of fixed effects and Z_1 be the random-effect design matrix that is used to account for repeated observations, the outcome model (1) is posed as

$$\left.\begin{array}{c}
\mathbf{Y}|\mathbf{W}(s)\sim \mathrm{EF}(\boldsymbol{\mu},\eta),\\
g(\boldsymbol{\mu})=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u},\\
\mathbf{u}\sim N\left\{\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{pmatrix},\begin{pmatrix}\sigma_{\mathbf{b}}^2\mathbf{I}_I & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{0} & \sigma_{\boldsymbol{g}_1}^2\mathbf{I}_{K_g-2} & & \mathbf{0}\\ \vdots & & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \dots & \sigma_{\boldsymbol{g}_L}^2\mathbf{I}_{K_g-2}\end{pmatrix}\right\},
\end{array}\right\} \tag{3}$$

where $\mathbf{X}=(1\;X\;(\mathbf{C}_1\mathbf{M}_1)^{[,1:2]}\dots(\mathbf{C}_K\mathbf{M}_K)^{[,1:2]})$ is the design matrix consisting of scalar covariates and fixed effects used to model the $\gamma_k(s)$, $\mathbf{Z}=(Z_1\;(\mathbf{C}_1\mathbf{M}_1)^{[,3:K_g]}\dots(\mathbf{C}_K\mathbf{M}_K)^{[,3:K_g]})$ is the design matrix consisting of subject-specific random effects and random effects used to model the coefficient functions, $\boldsymbol{\beta}=(\alpha,\beta,g_{10},g_{11},\dots,g_{K0},g_{K1})$ is the vector of fixed effect parameters and $\mathbf{u}=(\{b_i\}_{i=1}^I,\{g_{1m}\}_{m=3}^{K_g},\dots,\{g_{Km}\}_{m=3}^{K_g})$ is the vector of subject-specific random effects and the random effects used to model the $\gamma_k(s)$. The terms $a_k=\int_0^1 \mu_k(s)\,\gamma_k(s)\,\mathrm{d}s$ are incorporated in the overall model intercept α.

In this way longitudinal functional regression models can be flexibly estimated by using standard, yet carefully constructed, mixed effects models. Once again, we note that the random-effect structure is simple for expositional purposes only; complex combinations of random-effects and covariance structures can be implemented just as in standard generalized mixed models. Similarly to Crainiceanu and Goldsmith (2010), it is possible to model jointly the principal component loadings $\mathbf{c}_{ijk}$ and the outcome. This is important if there is substantial variability in the estimates of the $\mathbf{c}_{ijk}$ but may not be necessary if good estimates are available. We have assumed the same truncation lags K_w and K_g for each functional predictor and coefficient, but this assumption is easily relaxed. For the choice of truncation lags, we refer to Goldsmith *et al.* (2011a) and Ruppert (2002) and choose them large, subject to the identifiability constraint $K_w\geqslant K_g$; typically $K_w=K_g=30$ will suffice. Spline bases other than the truncated power series basis can (and, in the Bayesian software implementation, will) be used with appropriate changes to the specification of the random effects $\mathbf{u}$ that are used to model the $\gamma_k(s)$.

A related approach, which can be advantageous when the shape of $\gamma_k(s)$ is known, takes $\boldsymbol{\phi}_k(\cdot)$ as a collection of parametric functions, i.e. $\boldsymbol{\phi}_k(s)=\{\mathbf{1},t\}$ if $\gamma_k(s)$ is a linear function, and uses random effects to model the longitudinal structure of the data but not to induce smoothness on $\gamma_k(s)$. However, absent setting-specific knowledge of the shape of $\gamma_k(s)$ we typically advocate the more flexible penalized approach that was described above.

3. Simulations

We pursue a simulation study to test the effectiveness of our proposed method in estimating one or more functional coefficients in longitudinal regression models. We use two implementations, a likelihood-based approach and a Bayesian approach, and briefly note a few differences between the two. First, the likelihood-based approach estimates the matrix of principal component loadings **C** by using a Riemann sum approximation and then treats this matrix as fixed, whereas the Bayesian approach estimates **C** as in equation (2). Second, the Bayesian implementation uses a B-spline basis for the $\gamma_k(s)$ to improve mixing of the Markov chain Monte Carlo chains. A more detailed discussion of the two implementations is available in an on-line appendix. To ensure reproducibility, full code for the simulation exercise is available from `http://www.blackwellpublishing.com/rss`.

3.1. Single functional predictor

We first generate samples from the model

$$
\begin{aligned}
Y_{ij} &= b_i + \int_0^{10} W_{ij}(s)\gamma(s)\,\mathrm{d}s + \varepsilon_{ij}, \qquad i=1,\ldots,100,\\
W'_{ij}(s) &= W_{ij}(s) + \delta_{ij}(s),\\
W_{ij}(s) &= u_{ij1} + u_{ij2}s + \sum_{t=1}^{10}\left\{ v_{ijt1}\sin\left(\frac{\pi t}{5}s\right) + v_{ijt2}\cos\left(\frac{\pi t}{5}s\right)\right\}
\end{aligned}
$$

where $\varepsilon_{ij} \sim N(0,\sigma^2_{\mathbf{Y}})$, $\delta_{ij}(s) \sim N(0,\sigma^2_{\mathbf{W}})$, $b_i \sim N(0,\sigma^2_{\mathbf{b}})$, $u_{ij1} \sim \mathrm{Unif}(0,5)$, $u_{ij2} \sim N(1,0.2)$, v_{ijt1}, $v_{ijt2} \sim N(0,1/t^2)$ and $W'_{ij}(s)$ denotes the observed functional predictor for subject i at visit j. By construction the $W_{ij}(s)$ are a combination of a vertical shift, a slope and sine and cosine terms of various periods. The $W_{ij}(s)$ are observed on the dense grid $[s_g = g/10 : g=0,\ldots,100]$. We assume $I=100$ subjects, and we use two true coefficient functions in separate simulations: $\gamma_1(s) = 2\sin(\pi s/5)$ and $\gamma_2(s) = \surd s$. The first true coefficient function is selected to be an early principal component of the observed functions, and the second is an arbitrary smooth function. We take $J \in \{3,10\}$, $\sigma^2_{\mathbf{Y}} \in \{5,10\}$, $\sigma^2_{\mathbf{W}} \in \{0,0.5\}$ and $\sigma^2_{\mathbf{b}} \in \{5,50\}$, which give a total of 16 possible parameter combinations and two coefficient functions. For each of these combinations, we generate 100 data sets and fit model 1 by using both likelihood-based and Bayesian approaches. For the Bayesian implementation, we used chains of length 500 with the first 100 as burn-in; the estimated parameters are taken to be the posterior mean of the samples generated.

We calculate the mean-squared error of the estimated coefficient function $\hat{\gamma}(s)$ as

$$
\mathrm{MSE} = \sum_{g=0}^{100} \{\hat{\gamma}(s_g) - \gamma(s_g)\}^2.
$$

Table 1 provides the average mean-squared error (AMSE) by using both implementations taken over the 100 simulated data sets for $\sigma^2_{\mathbf{W}} = 0.5$, $\sigma^2_{\mathbf{b}} = 50$ and all possible combinations of J, $\sigma^2_{\mathbf{Y}}$ and the coefficient function; results for other values of $\sigma^2_{\mathbf{W}}$ and $\sigma^2_{\mathbf{b}}$ are quite similar to those presented and have been omitted. Thus, we see that the estimation of $\gamma(s)$ is very accurate regardless of the magnitude of the random-effect variance or, notably, the presence or absence of measurement error. As expected, there is a substantial decrease in AMSE when we observe 10 visits per subject compared with three visits per subject. Doubling the error variance on the outcome has the largest effect on the AMSE, but in many situations this is small. To provide context for Table 1, in Fig. 2, we show the estimates resulting in the median MSE for $J=3$, $\sigma^2_{\mathbf{W}} = 0$, $\sigma^2_{\mathbf{Y}} = 5$ and $\sigma^2_{\mathbf{b}} = 5$ under both coefficient functions.

We note that several differences between the results for the likelihood-based and Bayesian implementations are apparent for $J=3$. Particularly, note that the likelihood-based implemen-

Table 1. AMSE for the likelihood-based and Bayesian implementations with single functional predictors over 100 repetitions for $\sigma^2_{\mathbf{W}} = 0.5$ and $\sigma^2_{\mathbf{b}} = 50$, and all other possible parameter combinations and true coefficient functions†

J	*Method*	*AMSEs*			
		$\gamma_1(\cdot)$		$\gamma_2(\cdot)$	
		$\sigma^2_{\mathbf{Y}}=5$	$\sigma^2_{\mathbf{Y}}=10$	$\sigma^2_{\mathbf{Y}}=5$	$\sigma^2_{\mathbf{Y}}=10$
3	Likelihood	0.021	0.031	0.009	0.010
	Bayesian	0.017	0.036	0.014	0.019
10	Likelihood	0.009	0.013	0.006	0.007
	Bayesian	0.008	0.009	0.006	0.008

†Results are similar for other levels of $\sigma^2_{\mathbf{W}}$ and $\sigma^2_{\mathbf{b}}$.

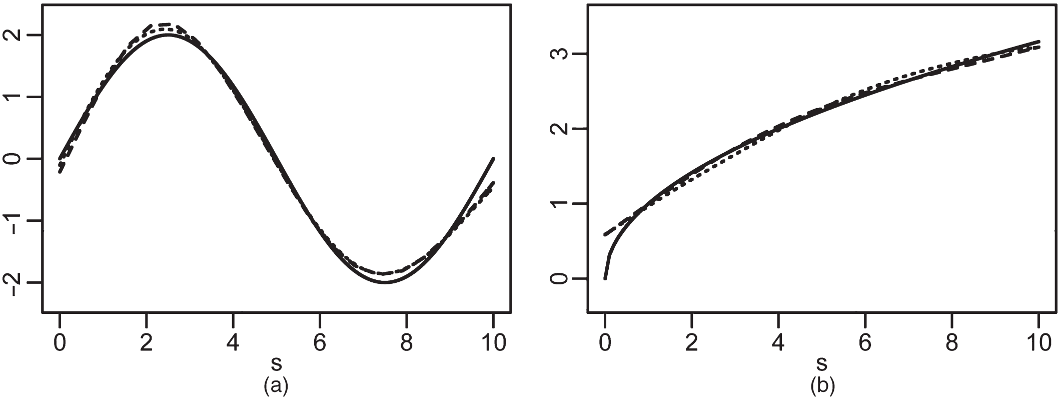

Fig. 2. For both true coefficient functions (———) and both implementations (– – –, likelihood based; - - - - - - -, Bayesian), the estimate $\hat{\gamma}(s)$ with median MSE: median MSEs for the likelihood-based and Bayesian approaches are respectively (a) 0.016 and 0.012 for $\gamma_1(s)$ and (b) 0.006 and 0.009 for $\gamma_2(s)$

tation generally has better performance for $\gamma_2(s)$. A possible reason for this is that the Bayesian implementation uses a smaller basis for both the functional predictors and the coefficient function and uses a *B*-spline basis for $\gamma(s)$, rather than a truncated power series. With fewer observations, the smaller *B*-spline basis may lack the flexibility to represent $\gamma_2(s)$ adequately. Though not shown, a second difference in the results for the different implementations is the slightly larger effect of measurement error on the AMSE for the Bayesian implementation. Recall that this approach jointly estimates the model parameters and the matrix of principal component loadings **C**. The added variability in estimating **C** in the presence of measurement error may lead to more variable estimation of the functional coefficient $\gamma(s)$. However, for $J=10$ these differences largely disappear.

Finally, in Fig. 3 we show the coverage probabilities for the 95% confidence and credible intervals produced by a subset of the simulations that were described above. We show the two extreme situations: first, we let $J=3$, $\sigma^2_{\mathbf{W}}=0$, $\sigma^2_{\mathbf{Y}}=5$ and $\sigma^2_{\mathbf{b}}=5$; second, we let $J=10$, $\sigma^2_{\mathbf{W}}=0.5$, $\sigma^2_{\mathbf{Y}}=10$ and $\sigma^2_{\mathbf{b}}=50$. For $\gamma_1(s)$, we generally see that the confidence intervals have coverage probabilities that are somewhat lower than the nominal level, whereas the credible intervals are

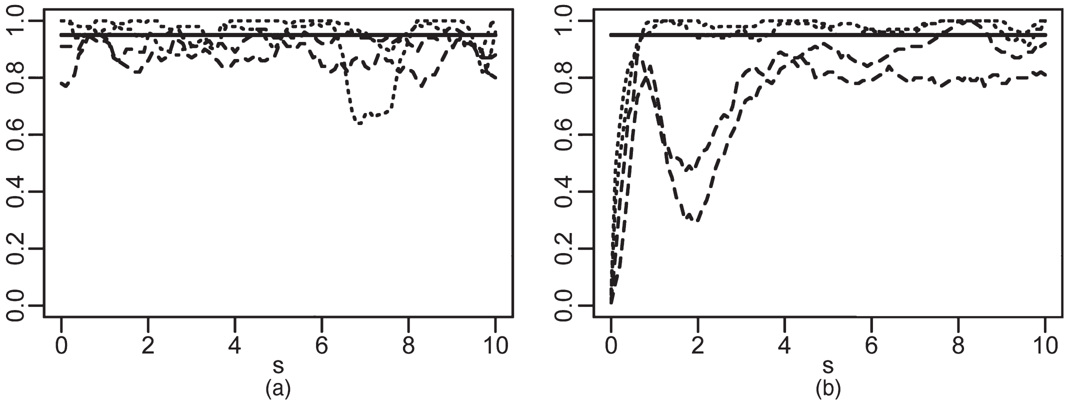

Fig. 3. Coverage probabilities for 95% confidence (– – –) and credible intervals (-------) for both true coefficient functions in simulations with single functional predictors: (a) $\gamma_1(s)$; (b) $\gamma_2(s)$

slightly conservative with the exception of one region. A more interesting situation is apparent for $\gamma_2(s)$. We note from Fig. 2 that both implementations tend to oversmooth the leftmost tail of the coefficient function; this is reflected in the very low coverage probabilities there. The Bayesian implementation recovers from this initial oversmoothing and has conservative credible intervals over the remainder of the domain, but the likelihood-based implementation has a second dip corresponding to a second region of oversmoothing before achieving coverage probabilities that are more similar to those for $\gamma_1(s)$. The oversmoothing of $\gamma_2(s)$ is most likely related to the use of a single parameter σ_g^2 to control smoothness across the domain of the coefficient function. In the case of $\gamma_2(s)$, the leftmost tail exhibits much more curvature than the remainder of the function, meaning that the single parameter may induce more smoothness than is accurate.

3.2. Multiple functional predictors

Next, we generate samples from the model

$$Y_{ij} = b_i + \int_0^{10} W_{ij1}(s)\gamma_1(s)\,\mathrm{d}s + \int_0^{10} W_{ij2}(s)\gamma_2(s)\,\mathrm{d}s + \varepsilon_{ij}, \qquad i = 1, \ldots, 100,$$

$$W'_{ijk}(s) = W_{ijk}(s) + \delta_{ijk}(s),$$

$$W_{ijk}(s) = u_{ijk1} + u_{ijk2}s + \sum_{t=1}^{10}\left\{ v_{ijtl1}\sin\left(\frac{\pi t}{5}s\right) + v_{ijtl2}\cos\left(\frac{\pi t}{5}s\right)\right\}$$

where $\varepsilon_{ijk} \sim N(0, \sigma_{\mathbf{Y}}^2)$, $\delta_{ijk}(s) \sim N(0, \sigma_{\mathbf{W}_k}^2)$, $b_i \sim N(0, \sigma_b^2)$, $u_{ijl1} \sim \text{Unif}(0, 5)$, $u_{ijl2} \sim N(1, 0.2)$, v_{ijtl1}, $v_{ijtl2} \sim N(0, 1/t^2)$ and $W'_{ijk}(s)$ denotes the kth observed functional predictor for subject i at visit j. Thus the $W_{ijk}(s)$ are independent functions constructed in the same manner as in Section 3.1. The $W_{ijk}(s)$ are observed on the dense grid $[s_g = g/10 : g = 0, \ldots, 100]$ and we set $I = 100$. Again we choose $\gamma_1(s) = 2\sin(\pi s/5)$ and $\gamma_2(s) = \sqrt{s}$ and we take $J \in \{3, 10\}$, $\sigma_{\mathbf{Y}}^2 \in \{5, 10\}$, $\sigma_{\mathbf{W}_k}^2 \in \{0, 0.5\}$ and $\sigma_{\mathbf{b}}^2 \in \{5, 50\}$. For each of these combinations, we generate 100 data sets and fit model 1 using both of the implementations that were described in Section 2. Owing to the added complexity of the model, we used chains of length 1000 with the first 500 as burn-in.

Table 2 provides the AMSEs resulting from both likelihood-based and Bayesian implementations of the longitudinal functional regression model with multiple functional predictors taken over the 100 simulated data sets; again, results for various combinations of $\sigma_{\mathbf{W}_k}^2$ and $\sigma_{\mathbf{b}}^2$ are similar and have been omitted. We see that the results in this setting are remarkably similar to

Table 2. AMSE for the likelihood-based and Bayesian implementations with multiple functional predictors over 100 repetitions for $\sigma^2_{\mathbf{W}_k}=0.5$ and $\sigma^2_{\mathbf{b}}=50$, and all other possible parameter combinations and true coefficient functions†

J	*Method*	*AMSEs*			
		$\gamma_1(\cdot)$		$\gamma_2(\cdot)$	
		$\sigma^2_{\mathbf{Y}}=5$	$\sigma^2_{\mathbf{Y}}=10$	$\sigma^2_{\mathbf{Y}}=5$	$\sigma^2_{\mathbf{Y}}=10$
3	Likelihood	0.019	0.028	0.008	0.011
	Bayesian	0.015	0.029	0.010	0.016
10	Likelihood	0.009	0.013	0.004	0.006
	Bayesian	0.008	0.009	0.004	0.006

†Results are similar for other levels of $\sigma^2_{\mathbf{W}_k}$ and $\sigma^2_{\mathbf{b}}$.

those considered in Section 3.1, despite the additional complexity of the model. Specifically, the AMSEs are negligibly affected and the comparisons between the likelihood-based and Bayesian implementations remain valid. Though not presented, figures examining the coverage probabilities of confidence and credible intervals are also largely unchanged.

4. Application to longitudinal diffusion tensor imaging regression

Recall that, in this study, 100 patients are scanned approximately once per year and undergo a collection of tests to assess cognitive and motor function; patients are seen between two and eight times, with a median of three visits per subject. Here we focus on the mean diffusivity profile of the *corpus callosum* tract and the parallel diffusivity profile of the right corticospinal tract as our functional predictors and the paced auditory serial addition test (PASAT), which is a commonly used examination of cognitive function affected by MS with scores ranging between 0 and 60, as our scalar outcome.

We begin by fitting models of the form

$$Y_{ij}=\alpha+X_{ij}\beta+b_i+\int_0^1 W_{ij}(s)\gamma(s)\,\mathrm{d}s+\varepsilon_{ij},\qquad b_i\sim N(0,\sigma^2_{\mathbf{b}}),\quad \varepsilon_{ij}\sim N(0,\sigma^2_{\mathbf{Y}})\qquad(4)$$

where Y_{ij} is the PASAT score for subject i at visit j, $W_{ij}(s)$ is the functional predictor for subject i at visit j and the variable $X_{ij}=I(j>1)$ is used to account for a learning effect that causes PASAT scores generally to rise between the first and second visit. Two such models are fitted: one using the mean diffusivity profile of the *corpus callosum* and another using the parallel diffusivity profile of the right corticospinal tract. Next, we fit additional models that include the subject-specific time since first visit as a fixed effect and random slopes on this variable. These models illustrate that more complex random-effect structures are possible within our proposed framework with a minimal increase in computation time and allow the treatment of irregularly timed observations. However, these models gave results that were indistinguishable from the random-intercept-only model and are not discussed further. The random-intercept models are fitted using both likelihood-based and Bayesian implementations of the method that was described in Section 2. Because our model uses a continuous outcome and identity link, the coefficient function has a marginal interpretation, i.e. the effect of the predictor function on the outcome (as mediated by the coefficient function) does not depend on a subject's random effect.

Table 3. Percentage of the variance in the PASAT outcome explained by a random-intercept-only model, a model without functional effects (labelled the 'non-functional model'), two models with single functional predictors (labelled 'single: mean diffusivity' and 'single: parallel diffusivity' for the models using the mean diffusivity and parallel diffusivity tract profiles respectively), and a model with multiple functional predictors (labelled 'multiple')

Parameter	*Results (%) for the following models:*				
	Random intercept only	*Non-functional model*	*Single: mean diffusivity*	*Single: parallel diffusivity*	*Multiple*
PVE	81.2	83.8	88.6	88.7	88.8

Estimates of the functional coefficient $\gamma(s)$, along with credible and confidence intervals, are presented in Figs 4(a) and 4(b). In both cases the credible and confidence intervals are quite different; we recall from our simulations that the credible intervals were conservative, whereas the confidence intervals were slightly below the nominal coverage. However, for the *corpus callosum* both intervals indicate that the first half of the tract has a significant influence on the PASAT outcome, although only the confidence intervals indicate significance for the region from 60 to 80. For the corticospinal tract, the region from 40 to 50 is indicated as having a significant effect on the outcome.

In Table 3 we show the percentage of the outcome variance explained, defined as $\text{PVE} = 100\{1 - \hat{\sigma}^2_{\mathbf{Y}}/\text{var}(Y_{ij})\}$ where $\hat{\sigma}^2_{\mathbf{Y}}$ is the estimated residual variance and $\text{var}(Y_{ij})$ is the overall outcome variance, in each of several models. Included in Table 3 are the longitudinal functional models that were described above, as well as a random-intercept-only model and a standard linear mixed model with visit indicator and the average mean diffusivity in the *corpus callosum* as a scalar predictor (rather than the functional tract profile). We note that the largest source of variation in the outcome is the subject-specific random effect, and that the functional regression models substantially outperform the standard linear mixed model. Further, the two functional predictors explain a similar amount of variability beyond the random-intercept-only model.

Next, we carry out an analysis that includes both the *corpus callosum* mean diffusivity profile and the corticospinal tract parallel diffusivity profile in a single model. The estimated functional coefficients in this model are given in Figs 4(c) and 4(d). The estimated coefficient function for the *corpus callosum* and accompanying confidence intervals are largely unchanged from regression using this profile as a single functional predictor, but the estimate for the corticospinal tract is quite different. Rather than a peak from 10 to 20 and a dip from 40 to 50, the coefficient for the corticospinal tract is near 0 over its entire range, and no regions appear particularly important in predicting the outcome. The model with multiple functional predictors explains only slightly more of the outcome variance than either model with a single predictor, indicating that the information that is contained in the two predictors is similar.

The collection of results from the regression models displayed in Table 3 is reminiscent of confounding in standard regression models: the effect of one variable largely disappears in the presence of a second variable. A heat map of the correlation matrix between the *corpus callosum* and corticospinal tract profiles, which is shown in Fig. 5, indicates that the region spanning locations 40–50 of the corticospinal tract have correlations between 0.5 and 0.6 with the *corpus callosum*. This region is similar to the region of interest when the corticospinal tract is used as a single predictor. The high correlations are probably due to the anatomical proximity of the

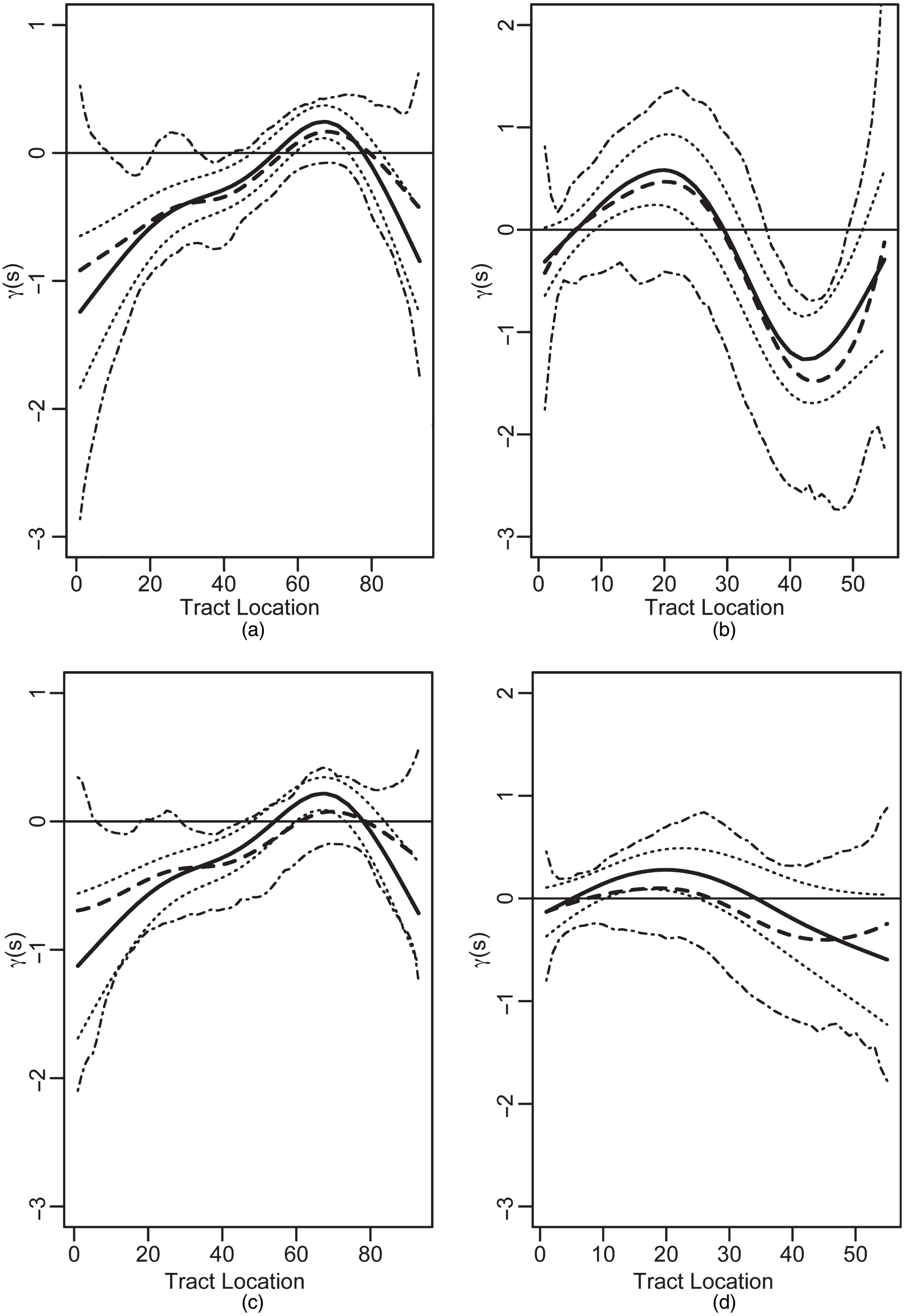

Fig. 4. Results of analyses with (a), (b) single and (c), (d) multiple functional predictors (———, estimated likelihood-based coefficient function; – – –, posterior mean; · · · · · · ·, 95% confidence interval; · - - - ·, 95% credible interval): (a), (c) estimated coefficient function for the median diffusivity profile of the *corpus callosum*; (b), (d) parallel diffusivity profile of the right corticospinal tract

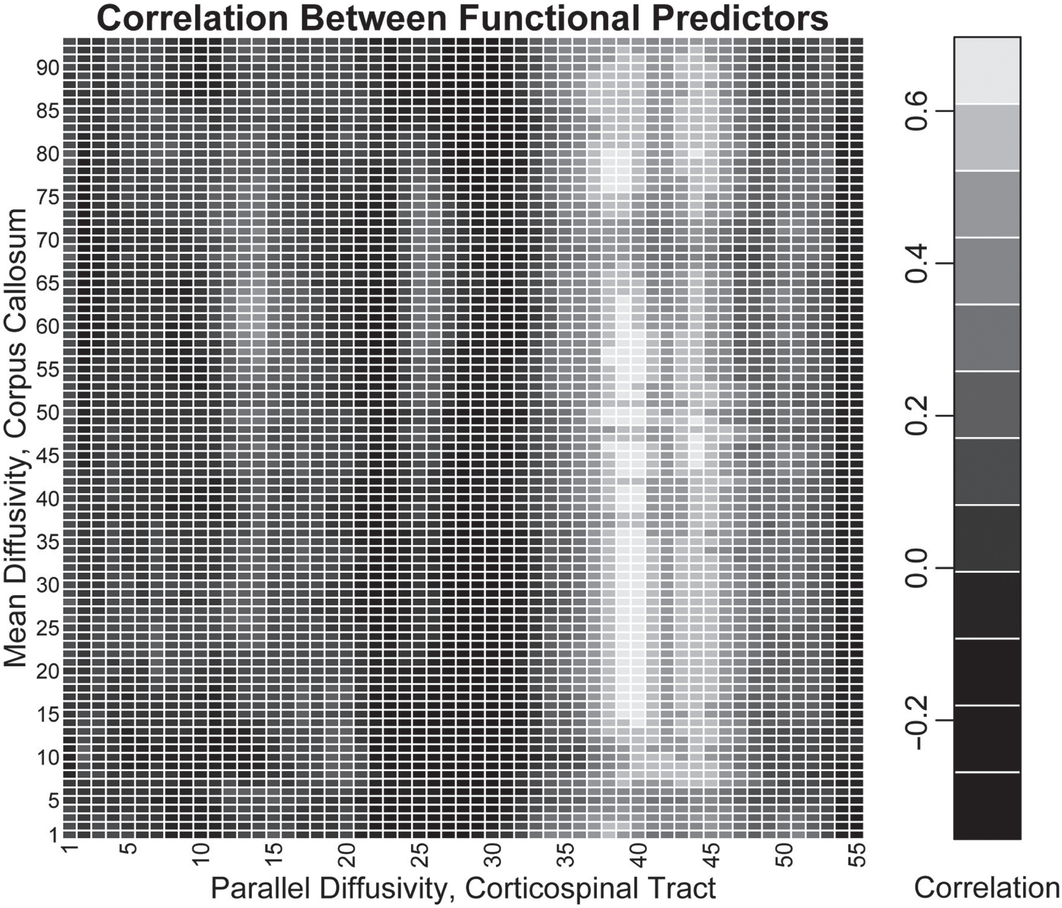

Fig. 5. Heat map of the correlation matrix between the mean diffusivity of the *corpus callosum* and the parallel diffusivity of the right corticospinal tract, the tract profiles used as functional predictors in the analyses with single and multiple functional predictors: tract locations are provided on the axes of the heat map

corpus callosum and corticospinal tracts in this region. Finally, the PASAT score is a measure of cognitive function and has been linked to degradation of the *corpus callosum* (Ozturk *et al.*, 2010). However, the corticospinal tract mediates movement and strength, not cognition, and therefore should not have a direct influence on the PASAT score. On the basis of these findings, it is therefore scientifically plausible that the corticospinal tract is not directly linked to cognitive function but is correlated through its correlation with the *corpus callosum*. However, the direct relationship between the *corpus callosum* and the PASAT score appears scientifically and statistically plausible.

5. Discussion

In this paper we are posed with a scientifically interesting and clinically important data set exploring the relationship between intracranial white matter tracts and cognitive disability in MS patients. Data of this type, in which functional predictors and scalar outcomes are observed longitudinally, will soon be regularly observed in the statistical community. To analyse these data properly, we proposed a longitudinal functional regression model which is broadly applicable. The results of our analysis indicate that the combination of functional data techniques with more traditional longitudinal regression models is a powerful tool to enhance our understanding of basic scientific questions.

The approach that we propose is appealing in that it casts difficult longitudinal functional regression problems in terms of the popular and well-known mixed model framework. The methods developed are

(a) very general, allowing for the functional coefficient to be fitted by using a flexible penalized approach or modelled parametrically,
(b) applicable whether the functional covariates are sparsely or densely sampled, or measured with error, and
(c) computationally efficient and tractable.

For the last point, we developed two implementations for this approach using common statistical software and demonstrated their effectiveness in a detailed simulation study; code for these implementations as well as a discussion of their relative merits are available in an on-line appendix and from `http://www.blackwellpublishing.com/rss`.

Future work may take several directions. Most broadly, there is a need for rigorous model selection techniques in the context of functional regression. In our analysis, we chose the model based on the scientifically plausible appearance of confounding, but the development of hypothesis tests that are suitable for this setting is necessary. Because in our model the flexibility in the coefficient function is based on a random-effect distribution, testing for a zero variance component can examine the null hypothesis that the effect of the functional predictor is constant over its domain. Methods exist to conduct this test, but their performance should be carefully studied in this setting. Alternatively, the use of lasso-based methods in similar settings allows penalized approaches to variable selection, although their feasibility in the longitudinal functional regression setting must be evaluated (Fan and James, 2011; Yi *et al.*, 2011). Related to the issue of variable selection, it is common for scientific applications to produce many correlated predictors for each subject. The inclusion of such predictors in a single model raises the issue of ‘concurvity’, which is a functional analogue of collinearity, and may result in unstable estimates or computational errors. Concurvity has been studied in the generalized additive model literature, but additional work is needed to address the problems that are caused by correlation in functional predictors (Ramsay *et al.*, 2003).

Other directions for future work include the possibility of basing longitudinal functional regression on functional decomposition techniques that are tailored to the structure of the data, which may give results that are simpler to implement or interpret. Penalized approaches that use cross-validation to impose smoothness, though computationally expensive, could provide a useful alternative to the mixed model approach that was presented here. Borrowing from the smoothing literature, the use of adaptive smoothing methods for the estimation of coefficient functions could alleviate the issue of localized oversmoothing or undersmoothing that was demonstrated in our simulations. Finally, functional regression techniques for data sets in which the predictor contains tens or hundreds of thousands of elements, either from extremely dense observation or because the predictor is a two- or three-dimensional image, will be needed as such data sets are collected and analysed.

Acknowledgements

The research of Goldsmith, Crainiceanu and Caffo was supported by award R01NS060910 from the National Institute of Neurological Disorders and Stroke. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Neurological Disorders and Stroke or the National Institutes of Health. This research was partially supported by the Intramural Research Program of the National Institute of Neurological Disorders and Stroke. Scans were funded by grants from the National Multiple Sclerosis Society and EMD Serono to Peter Calabresi, whose support we gratefully acknowledge.

References

Basser, P., Mattiello, J. and LeBihan, D. (1994) MR diffusion tensor spectroscopy and imaging. *Biophys. J.*, **66**, 259–267.

Basser, P., Pajevic, S., Pierpaoli, C. and Duda, J. (2000) In vivo fiber tractography using DT-MRI data. *Magn. Resnce Med.*, **44**, 625–632.

Brumback, B. and Rice, J. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Am. Statist. Ass.*, **93**, 961–976.

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

Cardot, H., Ferraty, F. and Sarda, P. (1999) Functional linear model. *Statist. Probab. Lett.*, **45**, 11–22.

Cardot, H., Ferraty, F. and Sarda, P. (2003) Spline estimators for the functional linear model. *Statist. Sin.*, **13**, 571–591.

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

Crainiceanu, C. M. and Goldsmith, J. (2010) Bayesian functional data analysis using WinBUGS. *J. Statist. Softwr.*, **32**, 1–33.

Crainiceanu, C. M., Staicu, A. M. and Di, C.-Z. (2009) Generalized multilevel functional regression. *J. Am. Statist. Ass.*, **104**, 1550–1561.

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

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S. and Punjabi, N. M. (2009) Multilevel functional principal component analysis. *Ann. Appl. Statist.*, **4**, 458–488.

Fan, Y. and James, G. M. (2011) Functional additive regression. To be published.

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

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2011a) Penalized functional regression. *J. Computnl Graph. Statist.*, to be published.

Goldsmith, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2011b) Penalized functional regression analysis of white-matter tract profiles in multiple sclerosis. *NeuroImage*, **57**, 431–439.

Greven, S., Crainiceanu, C. M., Caffo, B. and Reich, D. (2010) Longitudinal functional principal component analysis. *Electron. J. Statist.*, **4**, 1022–1054.

Guo, W. (2002) Functional mixed effects models. *Biometrics*, **58**, 121–128.

Hall, P., Müller, H.-G. and Wang, J. L. (2006) Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.*, **34**, 1493–1517.

James, G. M. (2002) Generalized linear models with functional predictors. *J. R. Statist. Soc.* B, **64**, 411–432.

James, G. M., Wang, J. and Zhu, J. (2009) Functional linear regression that's interpretable. *Ann. Statist.*, **37**, 2083–2108.

LeBihan, D., Mangin, J., Poupon, C. and Clark, C. (2001) Diffusion tensor imaging: concepts and applications. *J. Magn. Resnce Imgng*, **13**, 534–546.

Marx, B. D. and Eilers, P. H. C. (1999) Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics*, **41**, 1–13.

Mori, S. and Barker, P. (1999) Diffusion magnetic resonance imaging: its principle and applications. *Anat. Rec.*, **257**, 102–109.

Müller, H.-G. (2005) Functional modelling and classification of longitudinal data. *Scand. J. Statist.*, **32**, 223–240.

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

Ozturk, A., Smith, S., Gordon-Lipkin, E., Harrison, D., Shiee, N., Pham, D., Caffo, B., Calabresi, P. and Reich, D. (2010) MRI of the corpus callosum in multiple sclerosis: association with disability. *Mult. Scler.*, **16**, 166–177.

Ramsay, T. O., Burnett, R. T. and Krewski, D. (2003) The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*, **14**, 18–23.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*. New York: Springer.

Reiss, P. and Ogden, R. (2007) Functional principal component regression and functional partial least squares. *J. Am. Statist. Ass.*, **102**, 984–996.

Ruppert, D. (2002) Selecting the number of knots for penalized splines. *J. Computnl Graph. Statist.*, **11**, 735–757.

Staniswalis, J. and Lee, J. (1998) Nonparametric regression analysis of longitudinal data. *J. Am. Statist. Ass.*, **93**, 1403–1418.

Yao, F., Müller, H.-G., Clifford, A., Dueker, S., Follett, J., Lin, Y., Buchholz, B. and Vogel, J. (2003) Shrinkage estimation for functional principal component scores with application to the population. *Biometrics*, **59**, 676–685.

Yao, F., Müller, H.-G. and Wang, J. (2005) Functional linear regression analysis for longitudinal data. *Ann. Statist.*, **100**, 577–590.

Yi, G., Shi, J.-Q. and Choi, T. (2011) Penalized gaussian process regression and classification for high-dimensional nonlinear data. *Biometrics*, to be published.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Online Appendix to “Longitudinal Penalized Functional Regression for Cognitive Outcomes on Neuronal Tract Measurements”’.

Please note: Wiley–Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the author for correspondence for the article.