Biostatistics (2013), **14**, 3, *pp.* 447–461
doi:10.1093/biostatistics/kxs051
Advance Access publication on January 5, 2013

Longitudinal scalar-on-functions regression with application to tractography data

JAN GERTHEISS*

Department of Animal Sciences, Georg-August-Universität Göttingen, 37075 Göttingen, Germany
jgerthe@uni-goettingen.de

JEFF GOLDSMITH

Department of Biostatistics, Columbia University, New York, NY 10032, USA

CIPRIAN CRAINICEANU

Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA

SONJA GREVEN

Department of Statistics, Ludwig-Maximilians-Universität Munich, 80539 Munich, Germany

SUMMARY

We propose a class of estimation techniques for scalar-on-function regression where both outcomes and functional predictors may be observed at multiple visits. Our methods are motivated by a longitudinal brain diffusion tensor imaging tractography study. One of the study's primary goals is to evaluate the contemporaneous association between human function and brain imaging over time. The complexity of the study requires the development of methods that can simultaneously incorporate: (1) multiple functional (and scalar) regressors; (2) longitudinal outcome and predictor measurements per patient; (3) Gaussian or non-Gaussian outcomes; and (4) missing values within functional predictors. We propose two versions of a new method, longitudinal functional principal components regression (PCR). These methods extend the well-known functional PCR and allow for different effects of subject-specific trends in curves and of visit-specific deviations from that trend. The new methods are compared with existing approaches, and the most promising techniques are used for analyzing the tractography data.

Keywords: Diffusion tensor imaging; Functional principal components; Functional regression; Longitudinal functional principal components regression; Multiple sclerosis; Repeated measurements.

1. INTRODUCTION

Increasingly, longitudinal studies collect data, such as curves or images, that are functional in nature. Interest often centers on using these functional observations to predict longitudinal or time-invariant scalar

*To whom correspondence should be addressed.

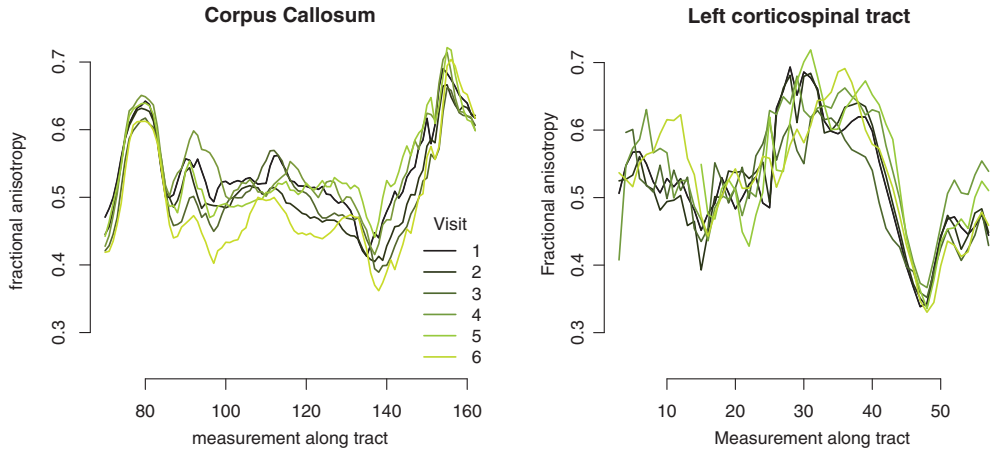

Fig. 1. FA along the corpus callosum and the left corticospinal tract for an MS patient, observed at 6 visits and measured at 93 and 55 sample points, respectively.

outcomes. To be specific, we are motivated by a neurological study on disease progression and corresponding changes in diffusion tensor images of the brain in multiple sclerosis (MS) patients. Interest lies in relating changes in neuronal tract properties extracted from the diffusion tensor images to disability scores measured at each visit, as well as in discriminating between MS patients and controls. Figure 1 displays the fractional anisotropy (FA) along the corpus callosum and the left corticospinal tracts for one of the MS patients observed at six different visits over a period of 4 years. In addition to FA, several other measurements of water diffusivity such as the magnetization transfer ratio (MTR) are available; measurements for other white matter tracts in the brain are also given. The study is one example of a rapidly increasing number of biomedical studies where, in contrast to simpler scalar-on-function regression, both outcomes and functional predictors are observed repeatedly over time, outcomes may be non-Gaussian, and there are multiple functional predictors. Moreover, any realistic method for such a problem will have to deal with additional non-functional covariates, such as age and sex, as well as partly missing or noisy functional predictors.

There exists a rich literature dedicated to scalar-on-function regression. For normally distributed outcomes, the *functional* linear model (Ramsay and Silverman, 2005) is implemented in the R package `fda` (Ramsay *and others*, 2012). It has been extended to non-Gaussian data by James (2002), Müller and Stadtmüller (2005), and James and Silverman (2005). Also P-splines as introduced by Marx and Eilers (1999) can be seen as a special case of this model. Ferraty and Vieu (2006) proposed methods for non-parametric functional regression and classification, Reiss and Ogden (2007, 2010) developed generalized functional principal components regression (FPCR) and partial least squares, and Goldsmith *and others* (2011) used the mixed-model methodology for fitting generalized functional linear models.

Despite these important advances, only longitudinal penalized functional regression (LPFR; Goldsmith *and others*, 2012) can currently handle regression where both the outcome and the functional data are measured at multiple visits. We observe data of the form $(Y_{ij}, X_{ij1}, \ldots, X_{ijp}, Z_{ij0}, \ldots, Z_{ijg}, W_{ij1}(s), \ldots, W_{ijq}(s))$, where Y_{ij} is the response for individual i at visit j $(i = 1, \ldots, n;\ j = 1, \ldots, n_i)$; $W_{ijm}(s)$ is the functional predictor over domain $\mathcal{D}_m$, $m = 1, \ldots, q$; $X_{ij} = (X_{ij1}, \ldots, X_{ijp})^\top$, and $Z_{ij} = (Z_{ij0}, \ldots, Z_{ijg})^\top$ are vectors of additional variables. For such data,

Goldsmith *and others* (2012) proposed the model

$$\mu_{ij} = h(\eta_{ij}) \quad \text{and} \quad \eta_{ij} = \alpha + \sum_{l=1}^{p} X_{ijl}\beta_l + \sum_{v=0}^{g} Z_{ijv}b_{vi} + \sum_{m=1}^{q} \int_{\mathcal{D}_m} W_{ijm}(s)\gamma_m(s)\,\mathrm{d}s, \tag{1.1}$$

with fixed effects $\beta_1, \ldots, \beta_p$, and independent and identically distributed vectors of random effects $(b_{0i}, \ldots, b_{gi})^\top = b_i \sim N(0, \Gamma)$. Function h is assumed to be a known link function. For given (non-functional and functional) covariates $X_{ij}, Z_{ij}, W_{ij1}, \ldots, W_{ijq}$ and random effects b_i, the distribution of Y_{ij} is assumed to be from an exponential family with conditional mean $\mu_{ij} = E(Y_{ij}|b_i, X_{ij}, Z_{ij}, W_{ij1}, \ldots, W_{ijq})$. For details about the generalized linear mixed model, see, e.g. McCulloch *and others* (2008). Model (1.1) can be estimated using the LPFR approach developed in Goldsmith *and others* (2012). Briefly, this method decomposes the functional predictors $W_{ijm}(s), m = 1, \ldots, q$, using functional principal components analysis (FPCA) ignoring the repeated and non-independent observation of curves within subjects i across visits j. Next, coefficient functions $\gamma_m(s)$ are expressed using a flexible spline basis. With these expansions, model (1.1) can be expressed in a mixed-model framework that induces smoothness in the coefficient functions and incorporates random effects b_i that account for correlation in the outcomes Y_{ij}. Several advantages of this approach are apparent. First, well-developed software can be used to fit such models either through the `refund` R package (Crainiceanu *and others*, 2012) designed with functional data analysis in mind or through the general mixed-model software in `mgcv` (Wood, 2006, 2011). The mixed-model framework also allows the construction of confidence intervals for estimated coefficient functions; see, e.g. Goldsmith *and others* (2011), Ruppert *and others* (2003), or Wood (2006) for details. Smoothing parameters that control the shape of the coefficient functions can be automatically estimated by maximum likelihood (ML) or restricted maximum likelihood (REML), and testing for constancy or linearity of the coefficient function is possible through testing whether the smoothing parameters are non-zero (Crainiceanu *and others*, 2005; Greven *and others*, 2008). FPCA expansion of functional predictors allows one to borrow strength across subjects in estimating basis functions; this is particularly useful when curves are partially unobserved.

However, LPFR does not directly address several inferential problems of interest. Foremost, it only uses subject-specific random intercepts to account for within-subject correlation of outcomes, but it does not explicitly account for the longitudinal structure of the functional predictor. In particular, the term $\int_{\mathcal{D}_m} W_{ijm}(s)\gamma_m(s)\,\mathrm{d}s$ in (1.1) does not separate the subject- and visit-level effects of the curve W_{ijm}. Separating these effects may be essential in scientific settings where one is interested in the association between individual components of variability and the outcome; a particular case of this problem occurs when one of the components is not actually associated with the outcome. Similarly, the FPCA decomposition used to expand functional predictors ignores the longitudinal structure of the observations and may miss important sources of variability.

Here, we propose to use the longitudinal FPCA (Greven *and others*, 2010) to extend the FPCR framework to the case when functional data are observed at multiple visits. Two versions will be developed. These approaches: (1) allow different subject- and visit-level effects on the outcome; (2) can be used when curves are observed with missings, or measured with error; (3) are applicable for both Gaussian and non-Gaussian outcomes; and (4) have freely available software implementations. The first point is highly relevant in applications where the interest will center on identifying a specific component of variability that is associated with the outcome. The second point will allow the efficient use of information and will avoid discarding predictors that exhibit missing data. For all computations, we used R (R Development Core Team, 2011), with code being provided in the appendix (see supplementary material available at *Biostatistics* online).

The remainder of the paper is organized as follows. In Section 2, we propose two versions of longitudinal FPCR (LFPCR). In Sections 3 and 4, we compare the different approaches via simulation studies

and analytically. The methods that performed well in the simulation studies are then used to analyze the tractography data (Section 5).

2. LONGITUDINAL FPCR

FPCR was proposed as an extension of principal components regression (PCR) (Massy, 1965; Frank and Friedman, 1993) to functional data; see Reiss and Ogden (2007) and references therein. In our case, functional predictors are not independent but are instead repeated observations on the same individuals. Therefore, we will present two possible ways of conducting LFPCR in this context.

2.1 *Longitudinal FPCA*

FPCA can be used for decomposing the variability in functional data. Since, however, some curves are obtained from the same individual, measurements are dependent. Hence, we use the functional random intercept and random slope model (Greven *and others*, 2010), where for subject i at visit j measurement $W_{ij}(s)$ at location $s \in \mathcal{D}$ is modeled as

$$W_{ij}(s) = \eta(s, T_{ij}) + B_{i,0}(s) + T_{ij} B_{i,1}(s) + U_{ij}(s) + \varepsilon_{ij}(s). \tag{2.1}$$

Time point T_{ij} indicates the time of visit j for subject i, and $\eta(s, T)$ is the overall smooth mean surface. The random processes $B_i(s) = \{B_{i,0}(s), B_{i,1}(s)\}$, $U_{ij}(s)$, and $\varepsilon_{ij}(s)$ are assumed to be mean zero, square-integrable, and mutually uncorrelated. The components $B_{i,0}(s)$ and $B_{i,1}(s)$ of $B_i(s)$ denote a functional random intercept and a random slope, respectively, capturing between-subject variation. $U_{ij}(s)$ is a visit-specific functional deviation from the subject-specific functional trend, capturing visit-to-visit functional variation on the same subject ("within-subject variation"). $\varepsilon_{ij}(s)$ is white noise error with variance ς^2, capturing random uncorrelated variation within each curve, cf. Greven *and others* (2010). Thus, model (2.1) allows one to decompose functional variation into three parts: subject-specific variation $B_i(s)$, visit-specific variation $U_{ij}(s)$, and measurement error. The Karhunen–Loève expansions of the processes $B_i(s)$ and $U_{ij}(s)$ using eigenfunctions $(\hat{\phi}_k^0, \hat{\phi}_k^1)$ and $\hat{\phi}_r^U$, respectively, are $B_{i,0}(s) = \sum_{k=1}^{\infty} \xi_{ik} \phi_k^0(s)$, $B_{i,1}(s) = \sum_{k=1}^{\infty} \xi_{ik} \phi_k^1(s)$, $U_{ij}(s) = \sum_{r=1}^{\infty} \zeta_{ijr} \phi_r^U(s)$, where the principal component scores $\xi_{ik} = \int_{\mathcal{D}} B_{i,0}(s) \phi_k^0(s)\, \mathrm{d}s + \int_{\mathcal{D}} B_{i,1}(s) \phi_k^1(s)\, \mathrm{d}s$ and $\zeta_{ijr} = \int_{\mathcal{D}} U_{ij}(s) \phi_r^U(s)\, \mathrm{d}s$ are uncorrelated random variables with mean zero and variances λ_k and ν_r, respectively. Longitudinal FPCA (LFPCA) estimates model (2.1) using a truncated version of the expansions above with N_B and N_U components, cf. Greven *and others* (2010). Thus, model (2.1) becomes $W_{ij}(s) \approx \eta(s, T_{ij}) + \sum_{k=1}^{N_B} \xi_{ik}(\phi_k^0(s) + T_{ij}\phi_k^1(s)) + \sum_{r=1}^{N_U} \zeta_{ijr} \phi_r^U(s) + \varepsilon_{ij}(s)$. For illustration, LFPCA of FA along the corticospinal tract is found in the appendix (see supplementary material available at *Biostatistics* online). For implementation details of LFPCA, refer to Greven *and others* (2010).

To choose the number of components N_B and N_U that are used to model the B and U processes, the proportion of explained variation can be used. Under some assumptions (in particular, for standardized visit times), total variation is given by $\int_{\mathcal{D}} \mathrm{Var}\{W_{ij}(s)\}\, \mathrm{d}s = \sum_{k=1}^{\infty} \lambda_k + \sum_{r=1}^{\infty} \nu_r + \varsigma^2$. So N_B and N_U may be chosen as the minimum numbers such that $\{\sum_{k=1}^{N_B} \hat{\lambda}_k + \sum_{r=1}^{N_U} \hat{\nu}_r + \hat{\varsigma}^2\} / \int_{\mathcal{D}} \widehat{\mathrm{Var}\{W_{ij}(s)\}}\, \mathrm{d}s \geqslant L$, where L is a pre-specified proportion of explained variance, such as $L = 0.90$ or 0.95, cf. Greven *and others* (2010).

Another important feature of LFPCA is that missing values in the curves are imputed automatically. Similarly to the simple FPCA-based approach (see Section 1), all available observations are used to estimate the principal component bases for all model components, and best linear unbiased prediction yields estimates of corresponding scores and curves. With LFPCA, however, curves are not simply pooled across

subjects, but the longitudinal structure of the data is taken into account, allowing the differential analysis of subject- and visit-level variability.

2.2 *Regression modeling using LFPCA scores*

The first and more intuitive of our LFPCA-based regression methods directly extends FPCR to the longitudinal setting. For modeling response Y_{ij} of subject i at visit j, we may use a PCR model where Y_{ij} is regressed on the scores ξ_{ik} and ζ_{ijr} from Section 2.1. To account for the repeated measures structure in Y_{ij}, we use a mixed model with subject-level random effects b_i. Furthermore, we note that scores in the LFPCA model (2.1) only refer to deviations from the mean surface $\eta(s, T_{ij})$. Therefore, we include a time-varying intercept $\int \varphi(s)\eta(s, T_{ij})\,\mathrm{d}s = \alpha(T_{ij})$, which can be estimated using penalized splines in the mixed models framework (see, e.g. Wood, 2011). Thus, our score-based LFPCR model is given by

$$\mu_{ij} = h(\eta_{ij}) \quad \text{and} \quad \eta_{ij} = \alpha(T_{ij}) + b_i + \sum_{l=1}^{p} \beta_l X_{ijl} + \sum_{k=1}^{N_B} \theta_k \xi_{ik} + \sum_{r=1}^{N_U} \delta_r \zeta_{ijr}, \tag{2.2}$$

where μ_{ij} denotes the conditional mean of Y_{ij} given the covariates and random effects. We assume $b_i \sim N(0, \tau^2)$ and conditionally independent Y_{ij} with a distribution from a simple exponential family. In addition to the scores ξ_k $(k = 1, \ldots, N_B)$ and ζ_r $(r = 1, \ldots, N_U)$, we specify fixed effects β_l, e.g. for age and sex. Potential random effects (beyond b_i) may be added as done in (1.1). Additional functional predictors would result in additional LFPCA scores and could thus be easily included. The model can also be simplified by focusing only on the scores from one level, as between-subject (ξ_{ik}) or within-subject (ζ_{ijr}) variation. For estimation of model parameters, analogously to traditional scalar PCR, scores obtained from LFPCA are used as scalar covariates in a non-functional regression model and coefficients can, for example, be estimated in the generalized additive mixed models framework (Wood, 2011) using corresponding software.

Owing to the construction of scores, model (2.2) can also be interpreted as a functional linear model where predictor and coefficient curves are expressed in the same orthonormal basis (see also Section 4). Similarly to more traditional FPCR, the number of retained scores can influence the regression results both quantitatively and qualitatively. N_B and N_U therefore act as implicit tuning parameters, controlling the amount of regularization induced. We choose N_B and N_U using the percent variance explained approach given in Section 2.1, and investigate and discuss how the quality of LFPCR is influenced by N_B and N_U in Sections 3 and 4.

2.3 *Functional regression using decomposed curves*

LFPCA provides estimates of the functional principal components $(\phi_k^0(s), \phi_k^1(s))$ and $\phi_r^U(s)$, $k = 1, \ldots, N_B$, $r = 1, \ldots, N_U$. Thus, between-subject variation $B_i(s, T_{ij}) = B_{i,0}(s) + T_{ij} B_{i,1}(s)$ over the domain of the functions, $\mathcal{D}$, and time T, as well as within-subject variation $U_{ij}(s)$ can be reconstructed using the scores. Here, $B_i(s, T_{ij})$ represents the systematic trend in subject i over time, while $U_{ij}(s)$ denotes visit-specific deviations from this trend. Both parts may be important as predictors. For example, $B_i(s, T_{ij})$ may be more relevant if $U_{ij}(s)$ constitutes mostly measurement error, while $U_{ij}(s)$ might be the more important component if curves that are *unusual* for this person are highly predictive for the outcome Y_{ij}.

Functional covariates $B_i(s, T_{ij})$ and $U_{ij}(s)$ can now be used in a functional regression model for Y_{ij}. Because $B_i(s, T_{ij})$ and $U_{ij}(s)$ only give deviations from the general trend $\eta(s, T_{ij})$ the intercept is allowed to vary over time. Similarly to model (1.1), we model the relationship between the functional covariates and

the outcome through coefficient functions expressed as penalized splines. Thus, our second, smoothing-based LFPCR model is

$$\mu_{ij} = h(\eta_{ij}) \quad \text{and} \quad \eta_{ij} = \alpha(T_{ij}) + b_i + \int_{\mathcal{D}} \gamma_B(s) B_i(s, T_{ij})\, \mathrm{d}s + \int_{\mathcal{D}} \gamma_U(s) U_{ij}(s)\, \mathrm{d}s, \tag{2.3}$$

with $B_i(s, T_{ij}) = B_{i,0}(s) + T_{ij} B_{i,1}(s)$ and $B_{i,0}(s) = \sum_{k=1}^{N_B} \xi_{ik} \phi_k^0(s)$, $B_{i,1}(s) = \sum_{k=1}^{N_B} \xi_{ik} \phi_k^1(s)$, $U_{ij}(s) = \sum_{r=1}^{N_U} \zeta_{ijr} \phi_r^U(s)$, and conditional mean μ_{ij}. As in (2.2), we assume a random intercept $b_i \sim N(0, \tau^2)$ and conditionally independent observations. Additional scalar covariates can be included as fixed (or random) effects. Additional functional predictors would be included as additional B and U processes resulting from LFPCA of these curves. After predictors $B_i(s, T_{ij})$ and $U_{ij}(s)$ are obtained using LFPCA, the coefficient functions γ_B and γ_U in model (2.3) are estimated using penalized spline expansions with tuning parameters estimated via ML or REML—for example, by using functions from `mgcv` (Wood, 2006, 2011). Thus, the implementation of model (2.3) is computationally analogous to that of (1.1), although because model (2.3) separates the systematic trend and visit-specific deviation it is interpretively unique.

Because in model (2.3) the covariates' influence on the response is modeled by smooth coefficient functions, it will be called *smooth* LFPCR from now on, to contrast it with *score* LFPCR (2.2) where scores are considered as scalar predictors in a (generalized) linear mixed model. In the smooth LFPCR approach, as in other penalized approaches to functional regression, smoothness in the coefficient functions is explicitly induced via a roughness penalty. Thus, the choice of N_B and N_U, which are implicit tuning parameters in the score LFPCR approach, is not particularly important for smooth LFPCR provided they are chosen large enough to capture important features in the predictors. Compared with score LFPCR, smooth LFPCR is particularly advantageous if the coefficient functions cannot be well expanded in the first few principal components (PCs) because large numbers of PCs can be included with much smaller risk of overfitting (see also Section 4).

3. SIMULATION STUDIES

In addition to LFPCR, we consider two versions of LPFR and four simple benchmark methods. The first LPFR approach uses predictor curves directly and penalized B-splines for fitting the coefficient functions (LPFR_B), while the second uses a truncated power basis for the coefficient functions. In LPFR_B, observations with missing values in the predictor curves (if any) are omitted; the second approach uses FPCA to impute missing values (as implemented in `refund`; Crainiceanu *and others* (2012)) and is henceforth referred to as LPFR_TRi. To investigate the effect of data imputation, we consider a simulation scenario with missing data (see below). With both LPFR_B and LPFR_TRi, we penalize deviations from a constant (c) and a linear function (l). The benchmarks are (1) a saturated model with massive overfitting (OF), i.e. $\hat{\mu}_{ij} = y_{ij}$, where y_{ij} denotes the observed response value for individual i at visit j, (2) a simple random intercept (RI) model without any covariates, (3) a random intercept model without covariates but including a smooth trend function $f(T_{ij})$ over time points T_{ij}, i.e. a model with time-effect (RI_te), and (4) a random intercept model (with smooth time-trend) where each functional covariate is simply averaged and thus included as a scalar predictor (RI_avfun). To investigate the effect of different choices of parameters N_B and N_U when applying LFPCR, we consider both score and smooth LFPCR with 90% and 95% of the functional covariates' variance being explained.

To evaluate the performance of each method, we use the observed mean squared error (MSE) $(1/n) \sum_{i,j} (\mu_{ij} - \hat{\mu}_{ij})^2$. This is the mean of squared differences between the true (conditional) mean μ_{ij} of individual i at visit j, given the covariates and random effects, and the corresponding estimated mean $\hat{\mu}_{ij}$, with n denoting the overall number of observations. The reason for considering this kind of MSE

rather than considering errors in estimated coefficient functions is that (a) it takes prediction of random effects into account and (b) the methods we consider pose distinct model structures from each other and from the data generating mechanism, preventing direct comparisons between model estimates, and true parameter values (see also Section 4).

In our first scenario, we consider single longitudinal predictor curves $W_{ij}(s)$ that are constructed according to the LFPCA model (2.1). For (ϕ_k^0, ϕ_k^1), we use an orthonormal sine/cosine basis, and for ϕ_k^U we take Legendre polynomials, as done by Greven *and others* (2010). Also, the visit times T_{ij} are simulated analogously to Greven *and others* (2010), such that the mean for each subject is zero, and increments $T_{ij} - T_{ij-1}$ are independent draws from $U[0, 1]$; then times are scaled to have unit variance. Scores ξ_{ik} and ζ_{ijr} are assumed to be normal with ξ-variances $\lambda_k = 0.5^{k-2}$, $k = 1, \ldots, 6$, and zero otherwise; for ζ-variances ν_r, we have $\nu_r = 0.5^r$, $r = 1, \ldots, 4$, and zero otherwise. For the measurement error variance, we assume $\varsigma^2 = 0.01$. Our design is unbalanced with on average four observations per individual $i = 1, \ldots, 100$. After generating the functional predictor curves, we simulate response values Y_{ij} according to model (1.1) with random intercept with variance $\tau^2 = 2$; Y_{ij} is assumed to be (conditionally) normal with variance $\sigma^2 = 2$. For the true coefficient function $\gamma(s)$, we consider (a) a nonlinear function having the shape of a Gamma-density with a rather sharp hump, (b) a linear, and (c) a constant function. The simulation scenario is designed such that the signal-to-noise ratio is similar to or smaller than the one found when analyzing the tractography data. Data generation, model estimation, and evaluation of the MSE is independently repeated 100 times. The resulting errors for scenario (a) are summarized in Figure 2 (top left). One can see that methods which assume the true underlying model structure (namely LPFR_B/TRi) perform best, but that LFPCR also performs quite well, in particular, smooth LFPCR. Results for (b) and (c) are similar (only shown in the appendix of supplementary material available at *Biostatistics* online). As expected, the performance of the smooth LFPCR is quite insensitive to the amount of variability explained by LFPCA, as long as one is generous with the variability explained. As noted in Section 2.3, the explicit regularization of coefficient functions imposed by the penalized spline expansion reduces the influence of the choice of N_B and N_U on parameter estimates.

In a second scenario, we keep σ^2 and τ^2 as before, but consider two functional predictors defined on $(0, s_{\max})$ and generated by $W_{ij\cdot}(s) = \frac{1}{100}(15 + \sum_{t=1}^{5} \theta_{ijt} \sin\{2\pi s(3 - \theta_{ijt})/s_{\max}\} - \vartheta_{ijt})$, where $\theta_{ijt} = \tilde{\theta}_{it} + \psi_{ijt}$ and $\vartheta_{ijt} = \tilde{\vartheta}_{it} + \upsilon_{ijt} \cdot \tilde{\theta}_{it}$, ψ_{ijt}, $\tilde{\vartheta}_{it}$, υ_{ijt} are independent random variables with $\tilde{\theta}_{it} \sim U[0, 4]$, $\tilde{\vartheta}_{it} \sim U[0, 2\pi]$, $\psi_{ijt} \sim U[-\frac{4}{5}, \frac{4}{5}]$, $\upsilon_{ijt} \sim U[-2\pi/5, 2\pi/5]$ for $W_{ij1}(s)$, and $\tilde{\theta}_{it} \sim U[0, 6]$, $\tilde{\vartheta}_{it} \sim U[0, 2\pi]$, $\psi_{ijt} \sim U[-\frac{6}{10}, \frac{6}{10}]$, $\upsilon_{ijt} \sim U[-2\pi/10, 2\pi/10]$ for $W_{ij2}(s)$. For $W_{ij1}(s)$ and $W_{ij2}(s)$, we have $s_{\max} = 50$ and $s_{\max} = 70$, respectively. Here, the LFPCA model (2.1) is not used for generating the functional predictors, but curves are directly simulated (similar to Tutz and Gertheiss (2010)). True $\gamma_1(s)$ and $\gamma_2(s)$ have Gamma-density-like shape (one with a sharper and one with a wider hump). The design is now balanced with five observations per subject $i = 1, \ldots, 100$. We add white noise measurement error with variances 0.008^2 and 0.004^2 to $W_{ij1}(s)$ and $W_{ij2}(s)$, respectively. Results (see appendix of supplementary material available at *Biostatistics* online) are similar like above: LPFR_B/TRi, which assume the correct model, perform best, but superiority over LFPCR is only moderate. When the data-generating process deviates from the simple model (1.1), however, LFPCR is distinctly superior to LPFR; see Figure 2 (top right, middle). Data generation for those and further scenarios is summarized in the following list: *Scenario* 3: the same specifications as in scenario 1, but now the ξ and ζ scores are used as predictors with all regression coefficients equal to 1; *Scenario* 4: the same specifications as in scenario 1, but now only the U process from (2.1) is used as functional predictor in a functional linear model with random intercept and with the true coefficient function having a Gamma-density-like shape again. This means, between-subject variation is only due to the random intercept. We use a Haar wavelet basis for (ϕ_k^0, ϕ_k^1), which yields predictor curves that should be harder to fit using LFPCA; *Scenario* 5: as in scenario 4, but now the U process is seen as an additional measurement error, and the only relevant functional predictor is the B process

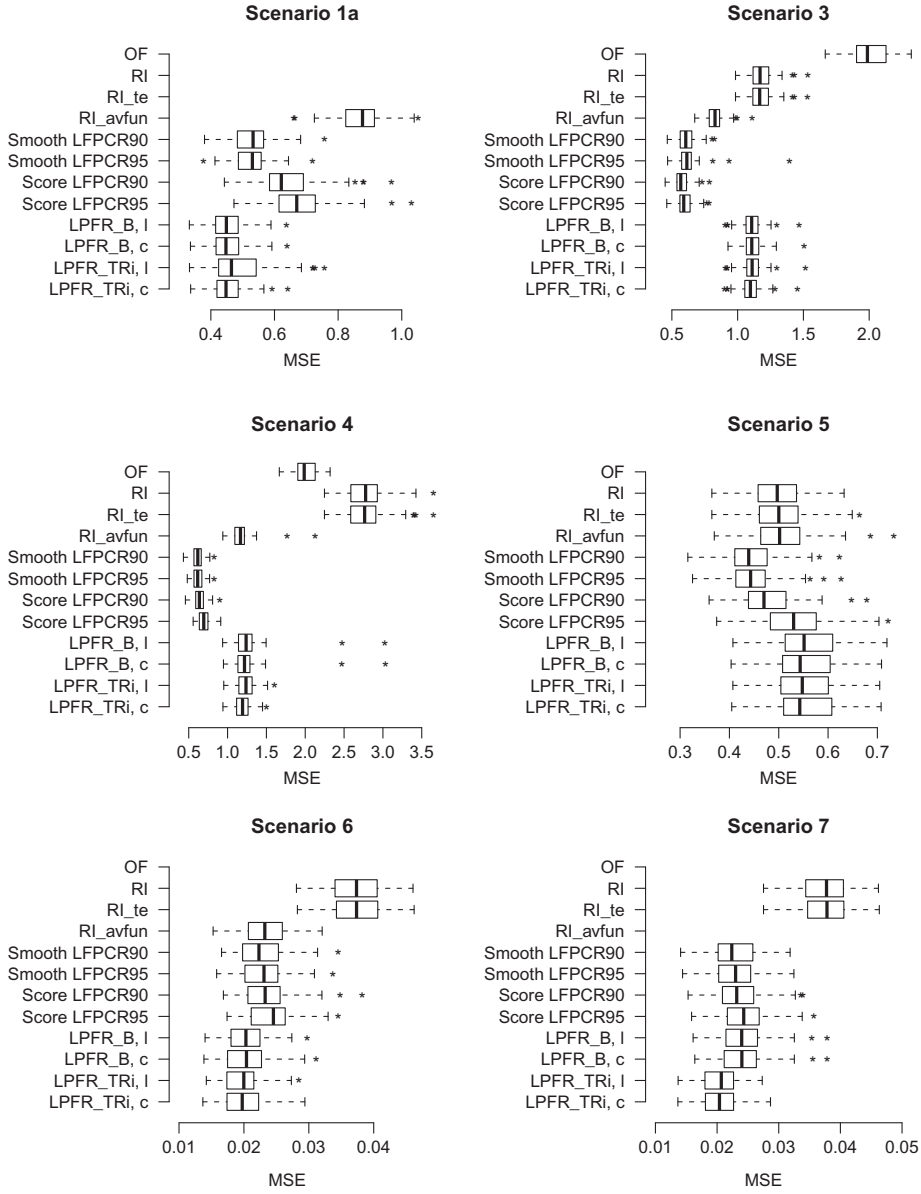

Fig. 2. Results of simulation scenarios 1a, 3–7 in terms of the (observed) MSE $(1/n)\sum_{i,j}(\mu_{ij} - \hat{\mu}_{ij})^2$ over 100 simulation runs. For some scenarios, results for the simple random intercept model (RI/RI_te/RI_avfun) and/or the OF model are not shown because error values were too extreme (in scenarios 1–5, OF of course produces errors around variance $\sigma^2 = 2$); for scenario 3 some outliers are not shown for RI, RI_te, RI_avfun, and LPFR_B.

$B_{i,0}(s) + T_{ij}B_{i,1}(s)$; *Scenario* 6: for generating the linear predictor the same specifications as in scenario 2 are used, but after employing the logistic function as the link binary outcomes are sampled; *Scenario* 7: as scenario 6, but in both sets of predictor curves 50 blocks of 2–4 missing observations are randomly selected.

It can be seen that in cases where the general functional trend $\eta(s, T)$ is irrelevant and only deviations from that trend are informative (scenarios 3–5), the LPFR model is inadequate. Furthermore, smooth

LFPCR seems to be superior to score LFPCR. In Section 4, we will have a closer look at connections between the different regression approaches. For the logit model (Figure 2, bottom), results are qualitatively similar to scenario 2. If data are missing in the predictor curves (Figure 2, bottom right), imputing missing values is superior to omitting curves with missings. The latter is done by LPFR_B, whereas LFPCR and LPFR_TRi use (L)FPCA-based imputation.

4. Comparing different approaches

Deeper insights into relations between the methods considered will help us understand some of the findings from the simulations. In Figure 2, we saw, for example, that smooth LFPCR tends to perform better than score LFPCR. The "functional part" in model (2.2) is

$$\sum_{k=1}^{N_B}\theta_k\xi_{ik}+\sum_{r=1}^{N_U}\delta_r\zeta_{ijr}=\sum_{k=1}^{N_B}\theta_k\left(\int_{\mathcal{D}}B_{i,0}(s)\phi_k^0(s)\,\mathrm{d}s+\int_{\mathcal{D}}B_{i,1}(s)\phi_k^1(s)\,\mathrm{d}s\right)+\sum_{r=1}^{N_U}\delta_r\int_{\mathcal{D}}U_{ij}(s)\phi_r^U(s)\,\mathrm{d}s$$
$$=\int_{\mathcal{D}}B_{i,0}(s)\sum_{k=1}^{N_B}\theta_k\phi_k^0(s)\,\mathrm{d}s+\int_{\mathcal{D}}B_{i,1}(s)\sum_{k=1}^{N_B}\theta_k\phi_k^1(s)\,\mathrm{d}s+\int_{\mathcal{D}}U_{ij}(s)\sum_{r=1}^{N_U}\delta_r\phi_r^U(s)\,\mathrm{d}s.$$

This is a functional linear model with predictors $B_{i,0}(s)$, $B_{i,1}(s)$, and $U_{ij}(s)$, and corresponding coefficient functions restricted to spaces spanned by the first eigenfunctions. These restrictions may explain problems in some situations. A more flexible approach would be to estimate these coefficient functions directly, which gives a third way to do LFPCR—a version that also uses B and U processes as predictors, but $B_{i,0}(s)$ and $B_{i,1}(s)$ separately. As $B_i(s,T_{ij})=B_{i,0}(s)+T_{ij}B_{i,1}(s)$ can be nicely interpreted as the systematic trend of subject i over time, however, we prefer the formulation in model (2.3). From the considerations above, it becomes clear that the latter approach and score LFPCR are neither equivalent nor is one a special case of the other.

Another important difference between smooth and score LFPCR is that in the latter case regularization is imposed by selecting the number of functional principal components. This number is tied to explaining the variability in the predictor curves and not to the parameter functions. If smooth LFPCR is applied, however, regularization through penalized splines is directly applied to the coefficient curves. This makes the approach robust to overfitting, even if N_B and N_U are very large. In contrast, for score LFPCR overfitting may become a problem if one chooses a proportion of explained variance that is too high or if the B or U process is not associated with the response (see simulation scenario 5).

Next, we investigate the differences between LPFR and smooth LFPCR seen in the simulation studies. Assume that model (1.1) and decomposition (2.1) hold. Then, after some straightforward algebra, we have $\int_{\mathcal{D}}W_{ij}(s)\gamma(s)\,\mathrm{d}s=\alpha(T_{ij})+\int_{\mathcal{D}}\gamma(s)B_i(s,T_{ij})\,\mathrm{d}s+\int_{\mathcal{D}}\gamma(s)U_{ij}(s)\,\mathrm{d}s+\tilde{\varepsilon}_{ij}$, where $\tilde{\varepsilon}_{ij}$ is noise with mean zero. If $W_{ij}(s)$ is a smooth curve without measurement error $\varepsilon_{ij}(s)$, the noise term $\tilde{\varepsilon}_{ij}$ disappears and the LPFR model (1.1) can be seen as a special case of LFPCR using B and U processes, where $\gamma_B(s)=\gamma_U(s)=\gamma(s)$ and $\alpha(T_{ij})$ has a specific form (see also the appendix of supplementary material available at *Biostatistics* online). Thus, if (1.1) holds, smooth LFPCR will also be an adequate modeling approach as long as (2.1) is a reasonable approximation to the (functional) data-generating process, as seen in simulation scenarios 1, 6, and 7. In contrast, if the LFPCR model is correct and the overall mean trend $\eta(s,T_{ij})$ is not relevant for the response, or if $\gamma_B(s)\neq\gamma_U(s)$, the functional linear part in (1.1) is not appropriate, as observed in scenarios 3–5. For illustration, we consider simulation scenarios 4 and 5, and compute the integrated squared error $\int_{\mathcal{D}}(\gamma_B(s)-\hat{\gamma}_B(s))^2\,\mathrm{d}s$, and $\int_{\mathcal{D}}(\gamma_U(s)-\hat{\gamma}_U(s))^2\,\mathrm{d}s$, respectively, for each simulation run. Table 1 gives the median values observed for LPFR (where $\gamma_B(s)=\gamma_U$ is assumed) and smooth LFPCR. In scenario 4, LPFR focusses on γ_B, causing high errors with respect to γ_U. As the

Table 1. *Median values of integrated squared errors* $\int_{\mathcal{D}}(\gamma_B(s) - \hat{\gamma}_B(s))^2\,\mathrm{d}s$ *and* $\int_{\mathcal{D}}(\gamma_U(s) - \hat{\gamma}_U(s))^2\,\mathrm{d}s$ *for LPFR and smooth LFPCR for simulation scenarios* 4 *and* 5

	Scenario 4		Scenario 5	
	γ_B	γ_U	γ_B	γ_U
LPFR_TRi,c	2.948	4.193	8.014	0.447
LPFR_TRi,l	2.743	4.740	8.559	0.287
Smooth LFPCR95	3.442	2.463	4.685	0.012
Smooth LFPCR90	3.448	2.462	4.251	0.010

Table 2. *Estimated fixed effects for scalar predictors when a generalized functional linear model with random intercept is directly fit to data with functional predictors FA and MTR along the corticospinal tract* (*LPFR, left*), *or when the model is fit to data with LFPCA* B *and* U *processes of the fractional covariates being used as functional predictors* (*LFPCR, right*)

	LPFR		LFPCR	
Variable	Estimated coefficient	p-value	Estimated coefficient	p-value
(Intercept)	4.3772	0.0000	2.8877	0.0000
Visit > 1	−0.0436	0.0111	−0.0598	0.0024
Sex	0.2377	0.0007	0.2049	0.0009
Age	0.0065	0.0165	0.0063	0.0098

U process is important this leads to large MSE, as seen in Table 2. Thus, LFPCR is more general than methods based on model (1.1), such as LPFR. If LPFR is correct, then the performance of LFPCR is only slightly affected. If specific LPFR assumptions are incorrect, then LFPCR outperforms LPFR.

5. APPLICATION TO THE TRACTOGRAPHY DATA

MS is a neurological disease that affects the central nervous system and in particular, damages white matter tracts in the brain through lesions, myelin loss, and axonal damage. Diffusion tensor imaging (DTI) is a magnetic resonance imaging technique that allows the extraction of information on individual tracts and thus allows a better understanding of damages in neuronal tracts and how these relate to disease progression. In our study, 176 MS patients were repeatedly scanned over time for an average of 1.27 years to follow disease progression, measured by disability scores such as the 9-hole peg test (*peg9*), and corresponding changes in DTI measurements. The *peg9* measures the time required to put nine pegs into nine holes and then remove them (cf. Cutter *and others*, 1999). Several summary indices including the FA and the MTR were extracted from the DTI images along several important tracts, including the corpus callosum, the corticospinal tract that is contralateral to the dominant hand, and the optic radiations tract. Our primary goal is to relate changes in disability to corresponding changes in tract profiles.

As seen in Section 3, LPFR performs well as long as model (1.1) corresponds to the true underlying regression structure. For the tractography data, it seems reasonable to assume that the functional covariates as a whole (for example, FA along a tract of interest) contain relevant information. Therefore, we start

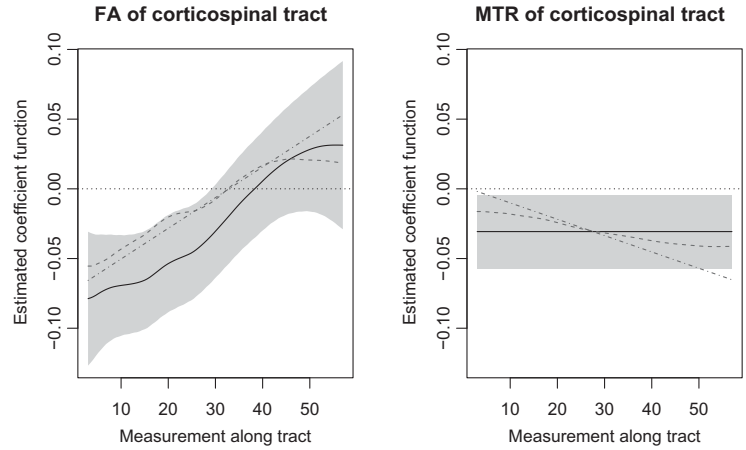

Fig. 3. Estimated coefficient functions when a generalized functional linear model with random intercept is fit to data with functional predictors FA and MTR of the corticospinal tract (and scalar predictors visit >1, sex, age). Considered are the results of a complete case analysis using LPFR with a B-spline basis for the coefficient functions and penalizing deviations from a constant line (solid), the same penalty but using a truncated power spline basis with knots at each observation point and imputing missing values using FPCA (dashed), and the `refund` implementation, where deviations from a linear function are penalized and missings are imputed (dashed/dotted). The shaded region corresponds to 90% pointwise confidence intervals as provided by `mgcv`.

with LPFR. After an initial analysis (see the appendix of supplementary material available at *Biostatistics* online), we only found a clear dependence between measures along the corticospinal tract and *peg9*, which is biologically plausible in that the corticospinal tract connects the motor cortex to the opposite side of the body, and therefore mediates motor signals. Hence, we consider FA and the MTR along the corticospinal tract as potential functional predictors for *peg9*. In addition, we consider scalar covariates sex and age, and a dummy variable indicating whether this is the patient's first visit or not. The latter is done to account for a potential learning effect with respect to the conducted test (see Goldsmith *and others*, 2012). Since the (conditional) distribution of the *peg9* scores is slightly skewed and scores are positive, we assume a Gamma distribution with log-link. Figure 3 shows the estimated coefficient functions. We tried different bases, penalties, and strategies for handling missing values as we did in Section 3. If curves with missing values are omitted, sample size is reduced by 38%. Since missings typically occur for technical reasons, however, it can be assumed that curves are missing at random, and hence a complete case analysis is reasonable. Apparently, there is a dependence between measures along the tract and *peg9*, with high values of FA in the first half of the tract resulting in lower disability scores. This makes sense because decreasing FA indicates disease progression (see Harrison *and others*, 2011). The solid and dashed lines for MTR, where deviations from a constant are penalized, indicate that using mean MTR here is a sufficient way of including anatomical information. Note that if deviations from linearity (dashed/dotted) are penalized, both coefficient curves are estimated to be non-constant, but linear. This, however, seems to us an artifact of the penalization. Results for the scalar covariates (complete case analysis) are given in Table 2 (left). Apparently, there is a learning effect after the first visit; and older and male patients tend to have higher *peg9* scores.

To investigate whether the assumption of equal effects for trend and deviation from the trend holds, we consider our new LFPCR approach. Following results from the simulation study, we use the better-performing smooth LFPCR. This means, we carry out LFPCA of FA and MTR along the corticospinal tract. The scores and functional principal components are then used to reconstruct the respective B and

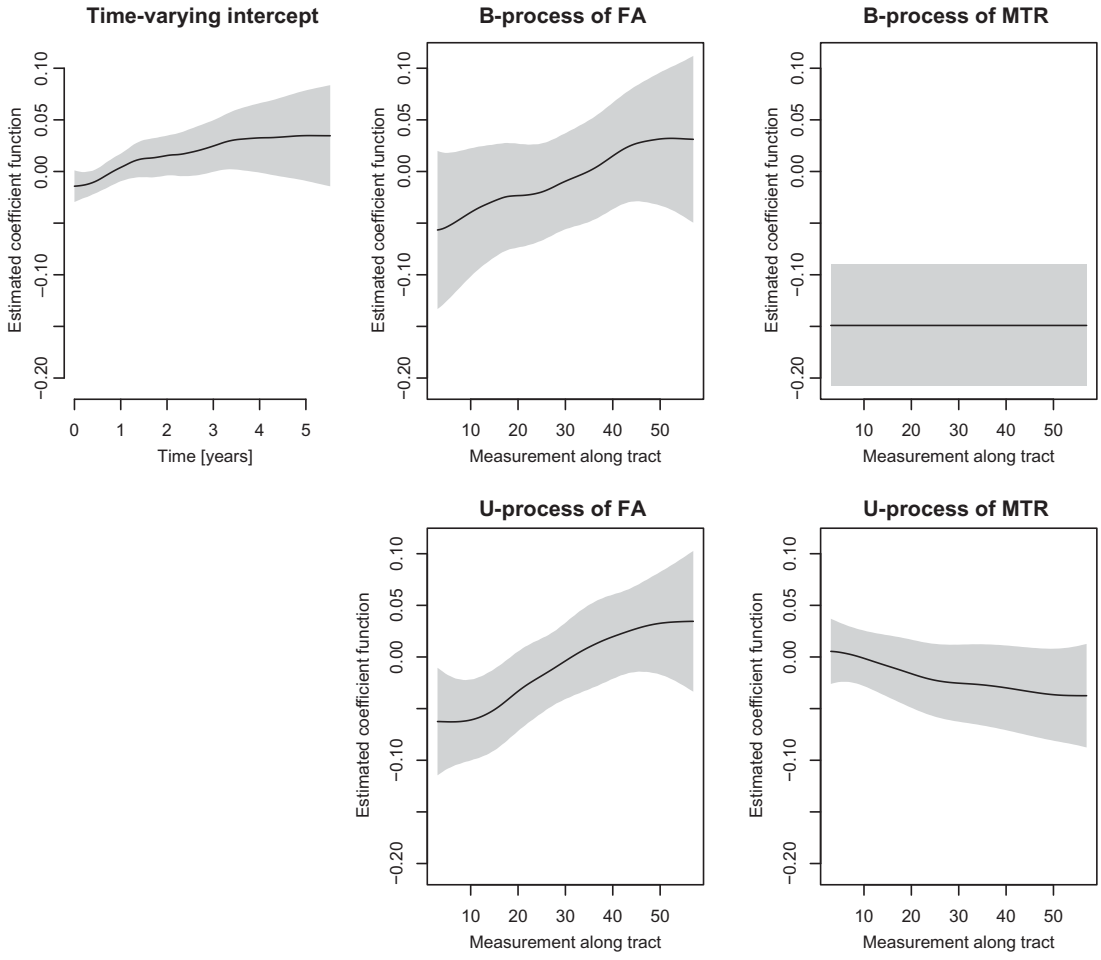

Fig. 4. Estimated coefficient functions when a smooth LFPCR model is fit to *peg9* scores with B and U processes of FA and MTR of the corticospinal tract being used as functional predictors, and with scalar predictors visit >1, sex, age.

U processes $B_i(s, T_{ij})$ and $U_{ij}(s)$ for each patient i at visit j. These curves are then used to build a generalized functional linear model with random intercept, and additional scalar predictors age, sex and visit >1 and time-varying intercept (see Section 2.3). With 90% variance being explained, we obtained $N_B = 15$ and $N_U = 9$ principal components for FA, and $N_B = N_U = 9$ for MTR. Although these values are rather large, due to explicit regularization, smooth LFPCR can be carried out without problems (see also Sections 3 and 4). Estimates for scalar covariates are given in Table 2 (right), estimated functions in Figure 4. Since missing values are already imputed when constructing B and U processes, differences between different fitting procedures for the regression model can be neglected, as only the bases differ. Effects of scalar covariates are estimated to be similar to estimates above (Table 2, left). The large difference between intercepts occurs because predictor curves are not centered, whereas the B and U processes are centered by construction (see Section 2.1). The time-varying intercept in Figure 4 (top left) indicates that the disability score increases over time, as expected in a diseased population. From the coefficient functions for both the B and U process of FA (Figure 4, middle), it follows that patients with

higher FA values than the "average patient" in the first half of the tract tend to have lower disability scores, which is in accordance with findings from above. Assuming equal coefficient functions for these two processes, as done with LPFR, seems reasonable. Estimated coefficient curves for MTR (Figure 4, right), however, indicate that mean MTR along the tract is highly informative with respect to *peg9* when, for example, different patients are compared (between-patients variation). Within-patient variation (the U process) of MTR, by contrast, seems to be less informative and may constitute mostly measurement error.

On the data at hand, LPFR and LFPCR predictions of *peg9* scores are very similar. With LPFR, the fitted function for MTR (see dashed line in Figure 3) seems to be a compromise between the coefficient functions of the B and U processes of MTR (see Figure 4, right). To generate a test set, we randomly selected one visit from each patient with four or more visits (yielding a test set of size 47), fit the regression models on the remaining training data, and predicted the test data. We repeated this procedure 100 times, and considered the test set deviances and the sum of squared errors to judge prediction accuracy. On average, test set deviances for LPFR and smooth LFPCR are virtually identical, but squared errors tend to be lower for smooth LFPCR (averaged squared error values for smooth LFPCR are by about 16% lower).

6. Summary and discussion

We presented and compared different tools for scalar-on-function regression that can be applied when observations are taken repeatedly over time. We proposed two novel versions of PCR for longitudinal functional data: score LFPCR and smooth LFPCR. While score LFPCR heavily depends on the number of principal components N_B and N_U that are used to describe the functional predictors, smooth LFPCR is robust against different choices of N_B and N_U as long as a large proportion, such as 90% or 95%, of variation in the functional covariates is explained. A possible way to alleviate problems with score LFPCR could be to use any variable selection approach. However, statistical inference becomes difficult in this context.

Furthermore, smooth LFPCR yields nice interpretations. For example, if deviations from subject-specific functional trends are just measurement error, irrelevant for the response variable, the corresponding coefficient function will be around zero. But as in a standard functional linear model, the coefficient function for the subject-specific trend will indicate the interesting regions in the signals' domain, and the functional shape of the influence on the response. LFPCR distinctly outperforms mixed models that use functional covariates directly when the overall trend in the functional predictors is not important for the response. On the other hand, it is competitive if the (generalized) functional linear model is (close to) the true model and the LFPCA model is a good approximation to the functional data generating process.

The presented LFPCR approaches can also be applied when only the functional predictors vary over time, but the response does not change from visit to visit. For example, consider the subjects' case status when it is to be discriminated between MS patients and controls. In a case like this, we focus on the subject-specific deviations from the overall trend, and use this processes as functional predictors in an adequate regression model; for example, in a logit model, if the aim of the analysis is binary classification (as case/control). As the trend in the functions might be an important predictor, advantages over taking just the functions' average can be expected. Such an analysis of the tractography data is provided in the appendix (see supplementary material available at *Biostatistics* online).

When LFPCA is carried out (R-code found at http://www.statistik.lmu.de/institut/ag/fda/research.html), all the proposed LFPCR methods result in a (generalized) additive mixed model with scalar or vector-valued random effects, scalar and functional fixed effects and potentially smooth effects of covariates such as time. Such models can, for example, be fit using R package `mgcv` (Wood, 2006, 2011).

Since the final regression model considered here is additive, it can be easily extended, for example, by 2D surfaces describing interactions of scalar predictors, modeling higher-dimensional functional predictors such as images or spatial effects. All such models can be implemented within the mixed-models framework in `mgcv` (Wood, 2006, 2011). Another possible extension would be to relax the linearity assumption with respect to visit times in the functional random slope model where LFPCA is based on. This may be done, for example, by allowing quadratic terms. For our data, however, variability of the B process over time is very low and the LFPCA model as used here is sufficient.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We thank Daniel Reich from the National Institute of Neurological Disorders and Stroke, the Department of Radiology and Neurology at Johns Hopkins Hospital, and the Department of Biostatistics at Johns Hopkins University for providing the data and related information. We thank two anonymous referees for a number of very helpful comments and suggestions. *Conflict of Interest:* None declared.

FUNDING

Gertheiss and Greven were supported by the German Research Foundation through the Emmy Noether grant GR3793/1-1. Goldsmith and Crainiceanu were supported by Award Number R01NS060910 from the National Institute of Neurological Disorders and Stroke. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of National Institute of Neurological Disorders and Stroke or the National Institute of Health.

REFERENCES

CRAINICEANU, C. M., REISS, P., GOLDSMITH, J., HUANG, L., HUO, L., SCHEIPL, F., GREVEN, S., HAREZLAK, J., KUNDU, M. G. AND ZHAO, Y. (2012). refund: Regression with Functional Data. R package version 0.1-6.

CRAINICEANU, C. M., RUPPERT, D., CLAESKENS, G. AND WAND, M. P. (2005). Exact likelihood ratio test for penalised splines. *Biometrika* **92**, 91–103.

CUTTER, G. R., BAIER, M. L., RUDICK, R. A., COOKFAIR, D. L., FISCHER, J. S., SYNDULKO, K., PETKAU, J., WEINSHENKER, B. G., ANTEL, J. P., CONFAVREUX, C. *and others*. (1999). Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain* **122**, 871–882.

FERRATY, F. AND VIEU, P. (2006). *Nonparametric Functional Data Analysis*. New York: Springer.

FRANK, I. E. AND FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109–148.

GOLDSMITH, J., BOBB, J., CRAINICEANU, C., CAFFO, B. AND REICH, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics* **20**, 830–851.

GOLDSMITH, J., CRAINICEANU, C., CAFFO, B. AND REICH, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society C (Applied Statistics)* **61**, 453–469.

Greven, S., Crainiceanu, C., Caffo, B. and Reich, D. (2010). Longitudinal functional principal components analysis. *Electronic Journal of Statistics* **4**, 1022–1054.

Greven, S., Crainiceanu, C., Küchenhoff, H. and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics* **17**, 870–891.

Harrison, D. M., Caffo, B. S., Shiee, N., Farrell, J. A. D., Bazin, P.-L., Farrell, S. K., Ratchford, J. N., Calabresi, P. A. and Reich, D. S. (2011). Longitudinal changes in diffusion tensor-based quantitative MRI in multiple sclerosis. *Neurology* **76**, 179–186.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society B* **64**, 411–432.

James, G. M. and Silverman, B. W. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association* **100**, 565–576.

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.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association* **60**, 234–256.

McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*, 2nd edition. Hoboken, NJ: Wiley.

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

R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Ramsay, J. O., Wickham, H., Graves, S. and Hooker, G. (2012). *fda: Functional Data Analysis*. R package version 2.3.2.

Reiss, P. T. and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association* **102**, 984–996.

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

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Tutz, G. and Gertheiss, J. (2010). Feature extraction in signal regression: a boosting technique for functional data regression. *Journal of Computational and Graphical Statistics* **19**, 154–174.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society B* **73**, 3–36.

[*Received January 20, 2012; revised July 12, 2012; accepted for publication October 30, 2012*]