Statistical Modelling 2017; **17(1–2)**: 1–35

A general framework for functional regression modelling

Sonja Greven[1] **and Fabian Scheipl**[1]
[1]Department of Statistics, Ludwig-Maximilians-Universität München, Germany

Abstract: Researchers are increasingly interested in regression models for functional data. This article discusses a comprehensive framework for additive (mixed) models for functional responses and/or functional covariates based on the guiding principle of reframing functional regression in terms of corresponding models for scalar data, allowing the adaptation of a large body of existing methods for these novel tasks. The framework encompasses many existing as well as new models. It includes regression for 'generalized' functional data, mean regression, quantile regression as well as generalized additive models for location, shape and scale (GAMLSS) for functional data. It admits many flexible linear, smooth or interaction terms of scalar and functional covariates as well as (functional) random effects and allows flexible choices of bases—particularly splines and functional principal components—and corresponding penalties for each term. It covers functional data observed on common (dense) or curve-specific (sparse) grids. Penalized-likelihood-based and gradient-boosting-based inference for these models are implemented in R packages refund and FDboost, respectively. We also discuss identifiability and computational complexity for the functional regression models covered. A running example on a longitudinal multiple sclerosis imaging study serves to illustrate the flexibility and utility of the proposed model class. Reproducible code for this case study is made available online.

Key words: functional additive mixed model, functional data, functional principal components, GAMLSS, gradient boosting, penalized splines

1 Introduction

1.1 Background and aims

Recent technological advances generate an increasing amount of functional data where each observation represents a curve or an image instead of a scalar or multivariate vector (Ramsay and Silverman, 2005; Horváth and Kokoszka, 2012). Functional data occur in medicine and biology, economics, chemistry and engineering as well as phonetics but are certainly not limited to these areas. Examples of technologies that generate functional data include imaging techniques, accelerometers, spectroscopy and spectrometry as well as any kind of measurement collected over time, data usually referred to as longitudinal. The term 'functional' data

Address for correspondence: Sonja Greven, Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80539 Munich, Germany.
E-mail: sonja.greven@stat.uni-muenchen.de

 10.1177/1471082X16681317

traditionally refers to data measured over an interval in the real numbers, although it is broader in meaning, for example also referring to functions on higher dimensional domains such as images over domains $\mathcal{T}$ in $\mathbb{R}^2$ or $\mathbb{R}^3$ or functions over manifolds. In this article, we will focus on functional data over a real interval $\mathcal{T}$, where curves could be observed on a dense grid common to all functions, with missings, or even on sparse irregular grids that are curve-specific. Sparse functional data commonly occur for longitudinal data that are viewed as functional data.

As functional data become more common, researchers are increasingly interested in relating functional variables to other variables of interest, that is in regression models for functional data. In addition, it becomes apparent that many complications well known from scalar data can and do also occur for functional data. Study designs or sampling strategies induce dependence structures between functions, for example, due to crossed designs, longitudinal or spatial settings. While more traditional functional data can be seen as realizations from stochastic processes that are often assumed to be Gaussian with some kind of smoothness assumption over the interval $\mathcal{T}$, there is a rising number of datasets where the observations consist of counts or binary quantities or follow skewed, bounded or otherwise non-normal distributions. It is thus also of interest to develop methods for 'generalized' functional data from non-Gaussian processes and/or to model other quantities of the conditional response distribution than (just) the mean.

In this article, we will focus on quite general, flexible models for regression with functional responses and/or covariates, with the aim of providing a similar amount of flexibility and modularity for functional data as the models that are presently available for scalar data—such as generalized additive mixed models (GAMMs), GAMLSS or (semi-parametric) quantile regression—and, in fact, strongly relying on recent advances in these areas. With this goal in mind, we will not provide a comprehensive review of available methods for functional regression (cf. Morris, 2015; Reiss et al., 2016; Wang et al., 2016), many of which are focused on one particular functional model at a time. We hope, instead, to provide a readable introduction to flexible functional regression within one overall consistent framework, also covering the implementation in R packages refund (Huang et al., 2016) and FDboost (Brockhaus and Rügamer, 2016). While the general framework we introduce, the notation we use and the estimation approaches we describe are largely based on the work of our own group and of our collaborators over the last few years, many of the particular functional regression models discussed in the literature (Morris, 2015; Reiss et al., 2016; Wang et al., 2016) can be seen as special cases of this framework. We point out connections and different approaches to estimation along the way, while keeping the focus on a unified set-up. We believe that having such a unified framework facilitates discussion, implementation and practical use of flexible functional regression models. Connecting their estimation to corresponding approaches for scalar data as we do here additionally ensures that recent and future advances in inference for such scalar regression models can be immediately used to expand the model class or improve inference for all functional models covered by this general framework. We are necessarily taking a somewhat subjective view coloured by the type of functional data and algorithms that we have worked on. Note that other approaches might

be better suited to different kinds of functional data including spiky (e.g., Morris et al., 2006) or truly big functional data (e.g., Zipunnikov et al., 2011; Reimherr and Nicolae, 2016).

1.2 Running example

As a running example, we will use a study on multiple sclerosis (MS) (Greven et al., 2010), which has been widely used as an example in the functional data literature. The dataset is available in the R package refund (Huang et al., 2016) and we can thus make our analysis fully reproducible in the code supplement provided for this article. This study followed 100 MS patients and 42 healthy controls longitudinally over time. At each of their up to eight visits (median number of visits: 2), subjects underwent a Diffusion Tensor Imaging (DTI) scan of their brain. MS patients additionally completed a Paced Auditory Serial Addition Test (PASAT) measuring abilities relating to information processing and attention, resulting in a scalar score. Fractional anisotropy (FA), which is related to directedness of water diffusion, was then extracted from the DTI scans along two major tracts in the brain—the corpus callosum (CCA) and the right corticospinal tract (RCST). FA is used as a proxy of demyelination, which acts as a marker of disease progression in MS since MS damages the myelin coating of the axons in the brain and thus impacts information transmission. FA values were averaged over slices of each tract, resulting in a scalar summary along one dimension of the tract. The procedure thus results in two functional variables for the two tracts, defined as functions of spatial distance along the tract. Figure 1 shows observed FA values for the two tracts we consider: The left panel shows CCA tracts, while the right panel shows centred and smoothed FA values along the RCST tracts. The coloured lines code for specific subjects (see the caption).

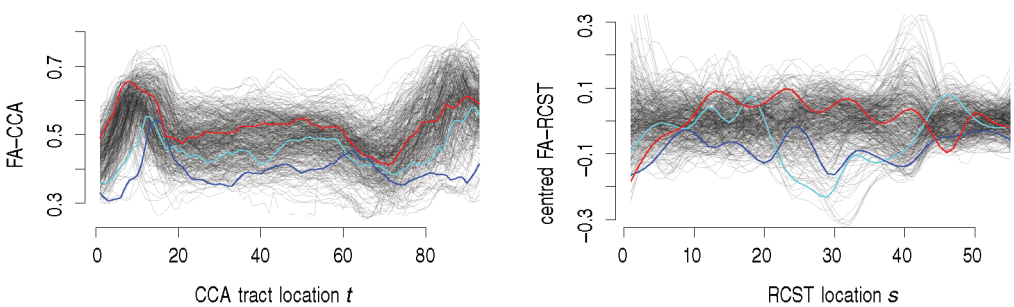

Figure 1 DTI data: Left panel shows observed FA values along the CCA tracts, right panel shows the centred and smoothed FA values along the RCST tracts. Blue and cyan lines show FA curves for selected MS patients with PASAT scores of 30 and 60, respectively. Red lines show curves for a control subject.

1.3 Functional regression models

The DTI data nicely illustrate that, depending on the question of interest, regression for functional data can occur in at least three flavours.

- If interest lies in quantifying the difference in FA–CCA profiles between cases and controls at the first visit, then we would use function-on-scalar regression (Reiss et al., 2010) for a functional response with a scalar covariate. A simple linear model would be

$$Y_i(t) = \beta_{v_i}(t) + E_i(t) + \varepsilon_{it}, \tag{1.1}$$

  where $Y_i(t)$ is the FA–CCA profile for subject i at the first visit at distance t along the tract, $\beta_{v_i}(t)$ represents a group-specific functional intercept with $v_i = 1[0]$ denoting subjects i belonging to the MS patients [controls], $E_i(t)$ a smooth residual and ε_{it} additional white noise error.
- If we are interested in whether the FA along the CCA is predictive of the PASAT score measured at the first visit, then we have a scalar response and a functional covariate, that is, scalar-on-function regression also known as signal regression (Marx and Eilers, 1999) or functional linear model (Cardot et al., 1999) if the effect is linear. In this case, a simple linear model for the first visits of the MS patients is

$$Y_i = \alpha + \int_{\mathcal{S}} x_i(s)\beta(s)ds + \varepsilon_i, \tag{1.2}$$

  with Y_i now the PASAT score for subject i and $x_i(s)$ denoting the FA–CCA profile observed at tract location s in $\mathcal{S}$, β an unknown weight or coefficient function and ε_i independent and identically distributed (i.i.d.) errors.
- If the focus lies on the relationship between the FA profiles along the two tracts, then we could think of a regression model with a functional response and a functional covariate (e.g., Ramsay and Silverman, 2005, Chapter 12), that is, function-on-function regression. A linear regression model in this case could be

$$Y_i(t) = \alpha(t) + \int_{\mathcal{S}} x_i(s)\beta(s,t)ds + E_i(t) + \varepsilon_{it}, \tag{1.3}$$

  with $Y_i(t)$ and $x_i(s)$ now referring to the CCA– and FA–RCST profiles observed at t in interval $\mathcal{T}$ and s in $\mathcal{S}$ respectively, $\beta(s,t)$ a bivariate coefficient quantifying the association between x at spatial location s with Y at spatial location t, and $E_i(t)$ and ε_{it} as in model (1.1).

Several extensions could also be considered. Models (1.1)–(1.3) all assume linear relationships between responses and covariates, which we might want to relax to more general smooth association structures. Also, the DTI data were collected longitudinally, and if we want to consider all observations simultaneously instead

of only the first visit, then we need to take into account the resulting correlation structure. Random effects are commonly used for scalar longitudinal observations and could be included in model (1.2), but for functional responses, we need functional analogs of random effects. We might want to modify model (1.3), for example, to

$$Y_{aw}(t) = \alpha(t) + \int_S x_{aw}(s)\beta(s,t)ds + B_a(t) + \varepsilon_{awt}, \tag{1.4}$$

where the double index now refers to the wth observation on the ath subject, $B_a(t)$ represents a subject-specific functional random intercept and the corresponding normality assumption of scalar random effects is replaced by a Gaussian process (GP) assumption. The smooth residuals E_i in models (1.1) and (1.3) are similarly modelled as curve-specific functional random effects. Finally, if the assumption of Gaussian responses does not fit the data well, we might want to change our models to, for example, quantile or generalized regression models.

1.4 Approaches

There is a large body of literature dealing with models like models (1.1)–(1.4) and further variants, and we refer to the recent comprehensive reviews by Morris (2015) and Reiss et al. (2016) for a full discussion. We can identify at least five general approaches to representing and modelling functional data and functional responses in particular, not mentioning extensive further work on estimation approaches specific to particular models. (For a recent overview on functional principal component (FPC)-based approaches, e.g, see Wang et al., 2016.)

The first general approach pre-smoothes each vector of observations along a function and then treats the resulting continuous curves as if they had been truly observed as objects in some function space (e.g., Ramsay and Silverman, 2005). This seems to be the historically first approach (Ramsay and Dalzell, 1991) and makes mathematical considerations somewhat easier. The downside in our view is that for the noisy observations common in many applications, the measurement error is not taken into account after the pre-smoothing step in the subsequent model. This approach also does not work (well) for sparse functional data and is not directly applicable to non-continuous data such as counts. Software implementations of this approach are available for R (package `fda`, Ramsay et al., 2014) and MATLAB (Ramsay et al., 2009).

Non-parametric methods for functional data have been proposed as non-parametric variants of this approach. Proposals for regression—mostly with just one functional covariate—and classification models for functional data in this framework are usually based on kernel methods and are distribution-free, so they are able to model highly non-linear, non-additive association structures and offer analysts great flexibility in specifying problem-specific semi-metrics for the kernels. However, extensions of these methods to multiple regression with several scalar and functional covariates or to 'generalized' functional data seem non-trivial. An overview for this theory and application examples are given in Ferraty and Vieu (2006); extensions

and a description of their implementation in the fda.usc R package are provided in Febrero-Bande and Oviedo de la Fuente (2012).

A third approach uses transformations of the response curves, usually projections into a coefficient space for a given set of basis functions, and subsequent multivariate modelling in this transformed space (e.g., Morris and Carroll, 2006; Morris et al., 2011). Such basis representations can be loss-less (e.g., wavelets) or lossy (e.g., truncated FPCs, explaining most of the variance in the data). This approach has computational advantages, as transformations can often be conducted with effort linear in the number of observation points and lossy transformations can be used for very high-dimensional data. Bases can also be tailored to the data at hand, for example, wavelets for spiky data or bases suitable for images. Disadvantages include that missings and curve-specific grids are difficult or impossible to handle in most of these approaches and that extensions to more general settings than mean regression for continuous data—for example, binary process data or quantile regression—are less than obvious. Fully Bayesian functional response regression methods based on a wavelet transformation are implemented in the `WFMM` software (Herrick, 2015).

A fourth approach is based on GP regression models (e.g., Shi et al., 2007; Shi and Choi, 2011) and directly models the observed functional data as realizations from such a GP with a covariance kernel from a known parametric family that typically incorporates covariate effects and linear effects of covariates on its mean function. Wang and Shi (2014) describe a generalization to non-Gaussian and dependent data where the underlying expectation is modelled using a latent GP. This approach is quite challenging computationally, as the optimization of the covariance parameters is a highly non-linear problem. A subset of this approach is implemented in the R package `GPFDA` (Shi and Cheng, 2014).

In the following, we will focus on a fifth approach, which directly models the observed data and expands all model terms in suitable basis expansions. To our knowledge, this approach was first described for the scalar-on-function case in Marx and Eilers (1999, 2005), with some early work given in Hastie and Mallows (1993). For functional responses, this approach is related to the literature on varying coefficient models (e.g., Hastie and Tibshirani, 1993; Reiss et al., 2010). Advantages in our opinion include facilitating accounting for all error sources in subsequent inference, allowing for the modelling of functional data observed on sparse or irregular curve-specific grids and going beyond mean regression for continuous functional data. In particular, this allows us to tackle quantile regression for functional data, generalized additive models for location, scale and shape (GAMLSS) as well as models for 'generalized' functional data such as data from binary or count processes. A further advantage not to be underestimated is that this approach reduces models for functional responses to models for scalar data, that is, models for the observed point values of each functional response. We thus avoid 'reinventing the wheel' and can take advantage of methods and algorithms for flexible regression models for scalar data that have been developed over the last decades. These include generalized additive (mixed) models (GA(M)Ms) (e.g., Eilers and Marx, 2002; Wood, 2006a; Schmid and Hothorn, 2008; Hothorn et al., 2016; Wood, 2016b), quantile regression (e.g., Koenker, 2005; Fenske et al., 2011)

or GAMLSS (e.g., Rigby and Stasinopoulos, 2005; Mayr et al., 2012). This also holds—at least to some extent—for inference in such models (e.g., Greven et al., 2008; Scheipl et al., 2008; Wood, 2013) and its transfer to functional regression (e.g., Staicu et al., 2014; Swihart et al., 2014; McLean et al., 2015).

Within the basis expansion approach, different basis functions such as FPCs, splines, wavelets or Fourier bases are conceivable and can be used. Different bases are well suited to different kinds of data: splines for smooth curves, wavelets for spiky functions and Fourier bases for periodic data. FPC bases are estimated from the data and work well if a large amount of variability is explained by relatively few modes of variation. These bases are commonly used with corresponding regularization penalties, such as smoothness penalties for splines or sparsity penalties for wavelet coefficients. In this work, we will particularly focus on spline bases with a smoothness penalty, assuming smoothness of the underlying functions over $\mathcal{T}$, and FPC bases, where the number of basis functions included in the model can be thought of as a discrete regularization parameter.

We introduce the proposed model class for flexible functional regression in Section 2 and discuss the specification of model terms in Section 3 and the estimation in Section 4. Section 5 covers identifiability in functional regression models and computational issues and we close with a discussion in Section 6. Code reproducing all analyses in this article using R packages refund and FDboost is available in an online supplement.

2 A general model formulation for functional data regression

2.1 A general functional regression model

We assume that we observe realizations from the following general regression model with functional responses and/or covariates (Brockhaus et al., 2015a,b, 2016b; Scheipl et al., 2015, 2016),

$$\boldsymbol{\xi}(Y|\boldsymbol{X}=\boldsymbol{x}) = h(\boldsymbol{x}) = \sum\nolimits_{j=1}^{J} h_j(\boldsymbol{x}). \tag{2.1}$$

Here, the response $Y \in \mathcal{Y}$ could be either scalar or (generalized) functional, with the space $\mathcal{Y}$ suitably chosen accordingly. To declutter notation, Y stands for the whole function $Y(t)$, $t \in \mathcal{T}$ and scalar responses are taken to correspond to the special case where $\mathcal{T}$ consists of a single value. Covariates $\boldsymbol{X} \in \mathcal{X}$ can include scalar and/or functional covariates and the space $\mathcal{X}$ is thus a suitable product space, with scalar covariates taking values in $\mathbb{R}$ and functional covariates over $\mathcal{S}$ assumed to be square integrable, that is, to lie in $L^2[\mathcal{S}]$.

The transformation function $\boldsymbol{\xi}$ for the conditional distribution of the response Y given the additive predictor indicates the feature of the conditional distribution that is modelled. (If h depends on latent processes, then model (2.1) also conditions on these processes and thus is a conditional model, analogous to the typical hierarchical formulation for mixed models.) The transformation $\boldsymbol{\xi}$ could correspond,

for example, to the (point-wise) expectation or median, a certain quantile, a link function composed with the expectation for, for example, count or binary process data, or a vector of several parameters such as mean and log-variance for GAMLSS for functional data.

This feature of the conditional response distribution is modelled in terms of an additive predictor $h(\boldsymbol{x}) = \sum_{j=1}^{J} h_j(\boldsymbol{x})$. (For GAMLSS, there are separate additive predictors for each component in the vector $\boldsymbol{\xi}$; see Brockhaus et al. (2015a, 2016a) for details. Each partial predictor $h_j(\boldsymbol{x})$ can depend on a subset of $\boldsymbol{x}$, thus also allowing for interaction terms that are functions of several covariates. Note that each $h_j(\boldsymbol{x})$ is also a real-valued function over $\mathcal{T}$ with values $h_j(\boldsymbol{x})(t)$. To obtain an identifiable model, certain constraints on the $h_j(\boldsymbol{x})$ are required which will be discussed in Section 5.1.

2.2 Examples

To give some intuition, consider again the models for the DTI data from the introduction. In the function-on-scalar model (1.1), $\boldsymbol{\xi} = \mathbb{E}$ is the expectation and we focus on mean regression $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \sum_{j=1}^{J} h_j(\boldsymbol{x})$. There are $J = 2$ partial predictors with $h_1(\boldsymbol{x}) = \beta_v$ depending on the scalar group indicator v and $h_2(\boldsymbol{x}) = E_i$ a smooth residual depending on the scalar curve indicator i. All model terms are functions over $\mathcal{T}$ spanning the length of the CCA tract. Results for this model are shown in the top panels of Figure 2.

If we are concerned about outlying values, then we could consider instead (point-wise) median regression by defining $\boldsymbol{\xi}$ to be the median. If we believe that measurement error might vary with the covariates and/or over the interval, then we can set $\boldsymbol{\xi} = (\mathbb{E}, \log \circ \mathrm{Var})^\top$ and model both conditional mean and conditional variance simultaneously as functions of $\boldsymbol{x}$ and t. Results for this model are shown in the middle row of panels in Figure 2. Note that this conditional variance function models heterogeneity of the variance of the white noise error term ε_{it}—autocorrelation and differences in the spread of the smooth underlying functions over $\mathcal{T}$ are modelled by the smooth residuals E_i. As FA values are, in fact, restricted to values in the $(0, 1)$ interval, a more suitable model than a Gaussian one might actually be a (point-wise) beta regression model. For this, we can take $\boldsymbol{\xi}$ to be $g \circ \mathbb{E}$, with g the logit link function. Results for this model are shown in the bottom panels of Figure 2. None of the three models is able to completely remove residual autocorrelation along t (Figure 2, right column), but the remaining autocorrelations are not very strong.

Extensions of the additive predictor also easily fit into this framework. The longitudinal function-on-function model (1.4), for instance, includes a functional random effect depending on the subject a. The function-on-function model (1.3) has $J = 3$ model terms depending on no covariates, on a functional covariate and on the curve indicator, respectively. Again, extensions are possible, for example, by allowing the effect of the functional covariate to be non-linear and changing the form of $h_2(\boldsymbol{x})(t)$ from $\int_{\mathcal{S}} x(s)\beta(s, t)ds$ to $\int_{\mathcal{S}} f(x(s), s, t)ds$ with a smooth unknown function f.

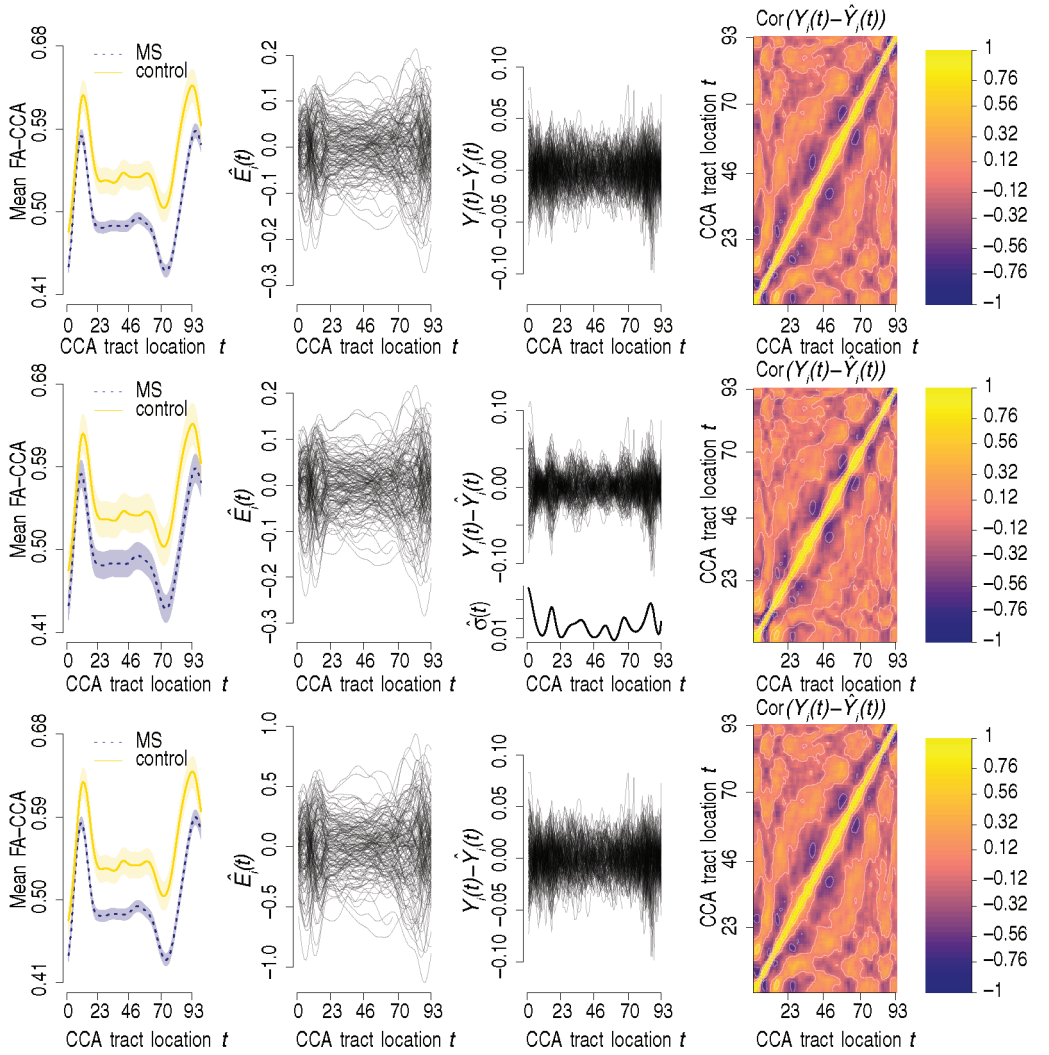

Figure 2 Results for (variants of) model (1.1) for a subset of the DTI data containing each subject's first visit. Top to bottom: Gaussian homoskedastic errors $\varepsilon_{it} \sim N(0, \sigma^2)$, Gaussian location-scale model with $\varepsilon_{it} \sim N(0, \sigma^2(t))$, beta regression model with logit link. Left to right: estimated group means $\hat{\beta}_0(t)$, $\hat{\beta}_1(t)$ with approximate point-wise 95% confidence intervals (25 cubic B-spline basis functions, first-order difference penalty), estimated smooth residuals $\hat{E}_i(t)$ (FPC basis with 8 FPCs, third row on latent logit scale), residuals $\hat{\varepsilon}_{it} = Y_i(t) - \hat{Y}_i(t)$, (second row with estimated variance function based on 25 cubic B-spline basis functions, first-order difference penalty), $t \in \mathcal{T}$, heatmap of correlation of residuals $\hat{\varepsilon}_{it}$ along t. Estimates produced with refund's `pffr` function (see Section 4.1).

The scalar-on-function model (1.2) corresponds to the special case of a scalar response with $\mathcal{T}$ collapsing to a single point and $h(\boldsymbol{x})$ taking values in $\mathbb{R}$ (see Section 3.5 for an application example and the right panel of Figure 3 for an example of a non-linear functional effect $\int_S f(x(s), s)ds$ in this context).

3 Specification of model terms

Model (2.1) introduces a general model class for regression with functional responses and/or covariates. For estimation of such models, we first discuss appropriate parameterizations using basis expansions for the model terms $h_j(\boldsymbol{x})$. We begin with some important special cases of $h_j(\boldsymbol{x})$ before embedding these into a more general framework, and then discuss the choice of bases.

We assume in the following that we observe realizations from model (2.1) indexed by $i = 1, \ldots, n$, where each response Y_i is measured on possibly curve-specific grid points $t_{i1}, \ldots, t_{iD_i}$ with $Y(t_{id})$ denoted by Y_{id}, $d = 1, \ldots, D_i$. Note that $D_i \equiv 1$ for the scalar response case.

3.1 Examples

3.1.1 Intercepts and scalar covariates

Consider again the models for the DTI data from the introduction. Recognizing that several of the model terms $h_j(\boldsymbol{x})$ in the functional response models can be seen as varying coefficient terms (Hastie and Tibshirani, 1993; Ruppert et al., 2003; Reiss et al., 2010), we can use well-known methods to approximate these model terms. For example, the smooth intercept curve in model (1.3) can be approximated as $h_j(\boldsymbol{x})(t) = \alpha(t) \approx \sum_{l=1}^{K_{Y_j}} \Phi_{Y_j,l}(t)\theta_{j,l}$, with h_j constant in the covariates $\boldsymbol{x}$, and the group effect in model (1.1) as $h_j(\boldsymbol{x})(t) = \beta_v(t) \approx \sum_{l=1}^{K_{Y_j}} \Phi_{Y_j,l}(t)((1-v)\theta_{j,1l} + v\theta_{j,2l})$, with h_j only depending on the scalar covariate v in the covariate set $\boldsymbol{x}$. For both, $j = 1$ in models (1.3) and (1.1), respectively, but the construction is general. We use a suitable basis $\{\Phi_{Y_j,l}, l = 1, \ldots, K_{Y_j}\}$—for example, splines—over $\mathcal{T}$ and unknown basis coefficients $\theta_{j,l}$ for all subjects in the first case or $\theta_{j,1l}$ and $\theta_{j,2l}$ for the control and MS groups, respectively, in the second case. The index Y indicates that $\Phi_{Y_j,l}$ is a basis over the response domain $\mathcal{T}$, while the index j corresponds to the model terms $h_j(\boldsymbol{x})$ that α and β_v represent. If age z had been available as a covariate, then we could have entered it into the model with a point-wise linear effect $h_j(\boldsymbol{x})(t) = z\gamma(t) \approx z\sum_{l=1}^{K_{Y_j}} \Phi_{Y_j,l}(t)\theta_{j,l}$. Alternatively, we could assume a smooth effect surface that is non-linear in z for each t using a tensor product basis $h_j(\boldsymbol{x})(t) = \gamma(z,t) \approx \sum_{k=1}^{K_{xj}} \sum_{l=1}^{K_{Y_j}} \Phi_{xj,k}(z)\Phi_{Y_j,l}(t)\theta_{j,kl}$ (De Boor, 1978; Eilers and Marx, 2003). The index x in $\Phi_{xj,k}$ indicates that $\Phi_{xj,k}$ is a basis depending on the respective covariate(s).

The intercept α in the scalar response model (1.2) can be seen as a special case of $\alpha(t)$, $t \in \mathcal{T}$, where we use a constant basis with one basis function, $\Phi_{Y_j,1}(t) \equiv 1$,

$K_{Yj} = 1$ and $\theta_{j,1} = \alpha$. We also take this approach for all covariate effects that are assumed to be constant in t in a functional response model.

3.1.2 *Functional random effects and smooth residuals*

Functional random effects B_a as in model (1.4) are assumed to be independent copies of a Gaussian random process. If $B_a(t)$ is assumed to be smooth in t for each level a, this GP is assumed to have a smooth covariance function. We can write $B_a(t) \approx \sum_{l=1}^{K_{Yj}} \Phi_{Yj,l}(t)\theta_{j,al} = \sum_{k=1}^{K_{xj}} \sum_{l=1}^{K_{Yj}} I(k = a)\Phi_{Yj,l}(t)\theta_{j,kl}$ (Scheipl et al., 2015) to obtain subject-specific functions using subject-specific coefficients $\theta_{j,al}$, where I is the indicator function selecting the relevant coefficients among all subjects' coefficients and K_{xj} is taken to be the number of subjects. More generally, for grouped data, a can be a grouping factor other than the subject and the B_a can be correlated over different levels of a. Smooth residuals as in models (1.1) and (1.3) correspond to the special case of functional random effects where the grouping variable is an identifier for each curve.

3.1.3 *Functional covariates*

For a linear functional covariate effect as in model (1.2), we can approximate the integral using numerical integration on the grid $s_1, \ldots, s_R$ of observation points in $\mathcal{S}$ (Wood, 2011)—here taken to be the same for all curves, although this could be generalized. The coefficient function can again be approximated (Marx and Eilers, 1999; Wood, 2011; Goldsmith et al., 2012) using a suitable basis, giving

$$h_j(\boldsymbol{x}) = \int_{\mathcal{S}} x(s)\beta(s)ds \approx \sum_{r=1}^{R} \Delta(s_r)x(s_r)\beta(s_r) \approx \sum_{r=1}^{R} \Delta(s_r)x(s_r) \sum_{k=1}^{K_{xj}} \Phi_{xj,k}(s_r)\theta_{j,k} \tag{3.1}$$

with suitable integration weights $\Delta(s_r)$.

For the functional response case, this is extended (Ivanescu et al., 2015) by simply replacing the basis for $\beta(s)$, $s \in \mathcal{S}$, by a suitable tensor product basis for $\beta(s, t)$, $s \in \mathcal{S}, t \in \mathcal{T}$,

$$h_j(\boldsymbol{x})(t) = \int_{\mathcal{S}} x(s)\beta(s, t)ds \approx \sum_{r=1}^{R} \Delta(s_r)x(s_r) \sum_{k=1}^{K_{xj}} \sum_{l=1}^{K_{Yj}} \Phi_{xj,k}(s_r)\Phi_{Yj,l}(t)\theta_{j,kl}.$$

In our DTI application, intervals $\mathcal{S}$ and $\mathcal{T}$ represent space and relating the covariate over the whole interval $\mathcal{S}$ to the response over the whole interval $\mathcal{T}$, thus, is of interest. In cases where functional responses and functional covariates are observed over the same time interval, it is often more meaningful to relate the response only to values of the covariate in the past (so-called historical models, Malfait and Ramsay, 2003). In this case, we can change the integration limits and integration weights accordingly

(Scheipl et al., 2015; Brockhaus et al., 2016b) and write

$$\begin{aligned} h_j(\boldsymbol{x})(t) &= \int_{\ell(t)}^{u(t)} x(s)\beta(s,t)ds \\ &\approx \sum_{r=1}^{R} I(\ell(t) \leq s_r \leq u(t))\Delta(s_r)x(s_r) \sum_{k=1}^{K_{xj}} \sum_{l=1}^{K_{Yj}} \Phi_{xj,k}(s_r)\Phi_{Yj,l}(t)\theta_{j,kl}, \end{aligned}$$

where $\ell(t)$ and $u(t)$ denote the lower and upper limits of integration. $[\ell(t), u(t)]$ may depend on t and could, for example, be $[0, t]$ or $[t - \delta, t]$ to allow for all previous covariate values or only values in a certain time window before the current time point to be associated with the response at a given t. The latter is directly related to distributed lags models for exposure-lag-response associations (e.g., Gasparrini et al., 2010; Obermeier et al., 2015). The limiting case of a concurrent effect $x(t)\beta(t)$ (e.g., Ramsay and Silverman, 2005) is achieved using $h_j(\boldsymbol{x})(t) = x(t)\beta(t) \approx x(t)\sum_{l=1}^{K_{Yj}} \Phi_{Yj,l}(t)\theta_{j,l}$.

Further extensions for the model terms contained in models (1.1) to (1.4)—such as non-linear effects of functional covariates or interaction terms—can be expressed similarly (see McLean et al., 2014; Brockhaus et al., 2015b; Fuchs et al., 2015; Scheipl et al., 2015; Usset et al., 2016). As is usual with such basis expansion approaches (e.g., Ruppert et al., 2003), regularization penalties can help in avoiding overfitting when large bases are used to provide flexibility in approximating underlying functions. We discuss such penalties in Sections 3.3 and 3.4.

3.2 General basis representation

In the examples discussed in Section 3.1, all the different model terms $h_j(\boldsymbol{x})$ have in common that we can express their basis representations in terms of one marginal basis parameterizing the effects of the covariates and another marginal basis parameterizing the effect's shape over $\mathcal{T}$. More generally, we write

$$h_j(\boldsymbol{x})(t) = (\boldsymbol{b}_{xj}(\boldsymbol{x})^\top \otimes \boldsymbol{b}_{Yj}(t)^\top)\boldsymbol{\theta}_j \tag{3.2}$$

for the terms $h_j(\boldsymbol{x})$ in model (2.1), with $\boldsymbol{b}_{xj}(\boldsymbol{x})$ the marginal basis vector for the covariate effect, $\boldsymbol{b}_{Yj}(t)$ the marginal basis vector over $\mathcal{T}$ and $\otimes$ denoting the Kronecker product. Importantly, the two marginal bases for each term can be chosen independently from one another and from those for the other terms, allowing for a flexible choice of bases appropriate for the problem at hand. $\boldsymbol{\theta}_j$ represents the unknown coefficient vector.

The examples from Section 3.1 fit into this general framework as follows: For effects $h_j(\boldsymbol{x})$ that vary over $\mathcal{T}$ like $h_j(\boldsymbol{x})(t) = \alpha(t)$, $\beta_v(t)$, $z\gamma(t)$, $\gamma(z, t)$ or $B_a(t)$, the basis vector $\boldsymbol{b}_{Yj}(t)$ parameterizing the effect's shape over $\mathcal{T}$ can contain any suitable basis $\Phi_{Yj,l}$, $l = 1, \ldots, K_{Yj}$, such as splines over $\mathcal{T}$, evaluated in t. For any model term in a

scalar response model or effects in a functional response setting that are assumed to be constant over t, $\boldsymbol{b}_{Yj}(t)$ is simply set to 1. The vector of coefficients is generally $\boldsymbol{\theta}_j = (\theta_{j,kl})_{k=1,\ldots,K_{xj};l=1,\ldots,K_{Yj}}$, where we drop the index l or k for simplicity in cases where $K_{Yj} = 1$ or $K_{xj} = 1$, respectively.

The basis vector $\boldsymbol{b}_{xj}(\boldsymbol{x})$ for the covariates depends on the specific covariate effect. For the global functional intercept, $\boldsymbol{b}_{xj}(\boldsymbol{x})$ is simply 1 as the effect is not associated with any covariate, that is,

$$h_j(\boldsymbol{x})(t) = \alpha(t) = \sum_{l=1}^{K_{Yj}} \Phi_{Yj,l}(t)\theta_{j,l} = \sum_{k=1}^{1}\sum_{l=1}^{K_{Yj}} 1 \cdot \Phi_{Yj,l}(t)\theta_{j,l} = (\boldsymbol{b}_{xj}(\boldsymbol{x})^\top \otimes \boldsymbol{b}_{Yj}(t)^\top)\boldsymbol{\theta}_j$$

with $\boldsymbol{b}_{Yj}(t)^\top = (\Phi_{Yj,1}(t), \ldots, \Phi_{Yj,K_{Yj}}(t))$ and $\boldsymbol{\theta}_j = (\theta_{j,l})_{l=1,\ldots,K_{Yj}}$. For the linear functional effect $z\gamma(t)$ of a scalar covariate z, the marginal basis vector in covariate direction is simply $\boldsymbol{b}_{xj}(\boldsymbol{x}) = z$. Similarly, $\boldsymbol{b}_{xj}(\boldsymbol{x})^\top = (1 - \nu, \nu)$ for $\beta_\nu(t)$, that is,

$$h_j(\boldsymbol{x})(t) = \beta_\nu(t) = \sum_{l=1}^{K_{Yj}}(1-\nu)\Phi_{Yj,l}(t)\theta_{j,1l} + \sum_{l=1}^{K_{Yj}}\nu\Phi_{Yj,l}(t)\theta_{j,2l} = (\boldsymbol{b}_{xj}(\boldsymbol{x})^\top \otimes \boldsymbol{b}_{Yj}(t)^\top)\boldsymbol{\theta}_j$$

with $\boldsymbol{b}_{Yj}(t)^\top = (\Phi_{Yj,1}(t), \ldots, \Phi_{Yj,K_{Yj}}(t))$ and $\boldsymbol{\theta}_j = (\theta_{j,kl})_{k=1,2;l=1,\ldots,K_{Yj}}$. For a smooth non-linear effect $h_j(\boldsymbol{x})(t) = \gamma(z,t)$, $\boldsymbol{b}_{xj}(\boldsymbol{x})$ contains spline-basis functions $\Phi_{xj,k}, k = 1, \ldots, K_{xj}$, evaluated in z. For a functional random effect, $h_j(\boldsymbol{x})(t) = B_a(t)$, with grouping variable a with K_{xj} levels, the basis vector $\boldsymbol{b}_{xj}(\boldsymbol{x})$ is an indicator vector of length K_{xj} for the levels of a.

For the linear functional term $h_j(\boldsymbol{x})(t) = \int_{\mathcal{S}} x(s)\beta(s,t)ds$, a spline-based approach would take $\boldsymbol{b}_{xj}(\boldsymbol{x}) = (\sum_{r=1}^{R} \Delta(s_r)x(s_r)\Phi_{xj,k}(s_r))_{k=1,\ldots,K_{xj}}$ and $\boldsymbol{b}_{Yj}(t) = (\Phi_{Yj,l}(t))_{l=1,\ldots,K_{Yj}}$, with spline basis functions $\Phi_{xj,k}$ and $\Phi_{Yj,l}$. Extensions such as interactions or non-linear terms can be similarly constructed (see, e.g., Brockhaus et al., 2015b; Scheipl et al., 2016).

In the construction in equation (3.2), we have implicitly assumed two points that are necessary for the Kronecker product construction to carry over to the design matrices. First, the grid over t must be the same for all curves, such that the basis over t evaluated at the grid points does not depend on the curve i. Second, the basis for the covariates needs to be the same for all values of t, eliminating any dependence of $\boldsymbol{b}_{xj}(\boldsymbol{x})$ on t. This is not fulfilled for functional historical terms $\int_{\ell(t)}^{u(t)} x(s)\beta(s,t)ds$ or concurrent effects $x(t)\beta(t)$, for example, where the basis vectors in covariate direction, $\boldsymbol{b}_{xj}(\boldsymbol{x},t)^\top = \left(\sum_{r=1}^{R} I(\ell(t) \le s_r \le u(t))\Delta(s_r)x(s_r)\Phi_{xj,k}(s_r)\right)_{k=1,\ldots,K_{xj}}$ respectively $\boldsymbol{b}_{xj}(\boldsymbol{x},t)^\top = x(t)$, depend on t. If either of these two requirements is not fulfilled, the Kronecker product construction in equation (3.2) is replaced by a row tensor product basis construction, $(\boldsymbol{b}_{xj}(\boldsymbol{x},t)^\top \odot \boldsymbol{b}_{Yj}(t)^\top)$, where $\boldsymbol{A} \odot \boldsymbol{B}$ for two $n \times p_A$

and $n \times p_B$ matrices is defined as $(\boldsymbol{A} \otimes \mathbf{1}_{p_B}^\top) \cdot (\mathbf{1}_{p_A}^\top \otimes \boldsymbol{B})$ with element-wise product $\cdot$ and $\mathbf{1}_p$ denoting a vector of ones of length p. In this construction, the marginal bases over $\boldsymbol{x}$ and t are first evaluated and then cross-multiplied for each curve and grid point separately (see Wood, 2006b; Brockhaus et al., 2015b; Scheipl et al., 2015) for details).

3.3 Regularization penalties

For regularization, we use penalties that are quadratic in the coefficient vector $\boldsymbol{\theta}_j$ containing all parameters for the jth effect $h_j(\boldsymbol{x})$. The general form for the quadratic penalty term is a Kronecker sum penalty (Eilers and Marx, 2003; Lang and Brezger, 2004; Wood, 2006a) $\boldsymbol{\theta}_j \boldsymbol{P}_j \boldsymbol{\theta}_j$ with

$$\boldsymbol{P}_j = \lambda_{xj} \boldsymbol{P}_{xj} \otimes \boldsymbol{I}_{K_{Yj}} + \lambda_{Yj} \boldsymbol{I}_{K_{xj}} \otimes \boldsymbol{P}_{Yj}, \tag{3.3}$$

where λ_{xj} and λ_{Yj} are smoothing parameters and $\boldsymbol{P}_{xj}$ and $\boldsymbol{P}_{Yj}$ are suitable marginal penalty matrices for the basis vectors $\boldsymbol{b}_{xj}(\boldsymbol{x})$ (or $\boldsymbol{b}_{xj}(\boldsymbol{x}, t)$) and $\boldsymbol{b}_{Yj}(t)$, respectively. For example, if we use B-splines in $\boldsymbol{b}_{Yj}(t)$ for effects such as $\alpha(t)$, $\beta_v(t)$, $z\gamma(t)$, then we can set $\boldsymbol{P}_{Yj}$ to a difference penalty matrix (P-splines; Eilers and Marx, 1996) and set the unneeded $\boldsymbol{P}_{xj}$ to 0; similarly, for $\int_{\mathcal{S}} x(s)\beta(s)ds$, where $\boldsymbol{P}_{xj}$ could be a difference penalty matrix if the $\Phi_{xj,k}$ in equation (3.1) are chosen as B-splines, and $\boldsymbol{P}_{Yj}$ is 0. For effects such as $\gamma(z,t)$, $\int_{\mathcal{S}} x(s)\beta(s,t)ds$ or $\int_{\ell(t)}^{u(t)} x(s)\beta(s,t)ds$, we typically need a smoothness penalty in both z/s and t directions and use corresponding Kronecker sum penalties as in equation (3.3). Likewise, for functional random effects $B_a(t)$, we use such a penalty with, for example, $\boldsymbol{P}_{xj} = \boldsymbol{I}_{K_{xj}}$ to reflect a normal distribution across independent factor levels (or some other precision matrix to define the dependence structure between the levels of a), and $\boldsymbol{P}_{Yj}$ corresponding to a smoothness penalty over t for each level of a.

Note the mathematical equivalence between the quadratic penalty (3.3) for $\boldsymbol{\theta}_j$ and a partially improper Gaussian prior $\boldsymbol{\theta}_j | \lambda_{xj}, \lambda_{Yj} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{P}_j^-)$ (Wood, 2006a, Chapter 4.8.1), where $\boldsymbol{A}^-$ denotes the (generalized) inverse of $\boldsymbol{A}$ as penalty matrices are typically only positive semi-definite. Consequently, the construction described here is equivalent to imposing a reduced-rank non-stationary GP prior on the model terms, with mean zero and covariance

$$\mathrm{Cov}(h_j(\boldsymbol{x})(t), h_j(\boldsymbol{x}')(t')) = (\boldsymbol{b}_{xj}(\boldsymbol{x}) \otimes \boldsymbol{b}_{Yj}(t))^\top \boldsymbol{P}_j^- (\boldsymbol{b}_{xj}(\boldsymbol{x}') \otimes \boldsymbol{b}_{Yj}(t')). \tag{3.4}$$

The choice of the marginal bases and penalties controls the prior covariance structure. From an empirical Bayesian perspective, inference for such terms can be performed based on established mixed models methods (see Section 4.1).

To achieve variable selection in scalar-on-function regression, Gertheiss et al. (2013b) use smoothness—sparseness penalties on coefficients for functional linear effects that are a combination of group LASSO-type penalties and the quadratic

roughness penalties described above. These might constitute a useful extension to the quadratic penalties in (3.3) we focus on here.

3.4 Choice of bases

Marginal basis vectors $\boldsymbol{b}_{xj}(\boldsymbol{x})$ and $\boldsymbol{b}_{Yj}(t)$ in equation (3.2) can be chosen freely as appropriate for the given modelling task. For functional responses, different bases in $\boldsymbol{b}_{Yj}(t)$ for representing the model terms in t direction have different properties and are chosen accordingly.

Pre-defined bases include splines and wavelets. Spline bases such as B-splines or truncated powers are commonly used to represent smooth terms, when the functional responses are smooth up to i.i.d. error. They are often used together with a quadratic smoothing penalty such as difference-based (P-splines; Eilers and Marx, 1996) or derivative-based (O'Sullivan penalized splines; O'Sullivan, 1986; Wand and Ormerod, 2008) penalty terms for B-splines and a penalization of the truncated polynomial terms for the truncated power series basis (e.g., Ruppert et al., 2003).

Wavelets are well-suited to spiky functional data or functional data with local features (for their use with functional data, see, e.g., Morris and Carroll, 2006). Due to their multiscale representation, and different from equation (3.3), they are commonly used with thresholding or L1-type penalties for coefficients to encourage denoising by shrinking small coefficients to zero.

FPC bases, by contrast, are estimated from the data. In model (1.1), for example, E_i are independent functional random intercepts and thus independent copies of a stochastic process assumed to have a smooth covariance, $C^E(s,t) = \mathrm{Cov}(E_i(s), E_i(t))$. Using Mercer's theorem and the Karhunen–Loève expansion (Mercer, 1909; Loève, 1945; Karhunen, 1947), we can write

$$E_i(t) = \sum_{l=1}^{\infty} \theta_{j,il}\phi_l^E(t),$$

where the $\phi_l^E, l \in \mathbb{N}$ are orthonormal eigenfunctions of the covariance operator associated with C^E for eigenvalues $\kappa_1^E \geq \kappa_2^E \geq \cdots \geq 0$, $\theta_{j,il}$ are uncorrelated random variables with mean zero and variance κ_l^E; $\theta_{j,il}$ are independent $\mathcal{N}(0, \kappa_l^E)$ variables if E_i is a GP. FPCs, that is, estimated eigenfunctions $\hat{\phi}_l^E$, can be used as a basis for E_i in $\boldsymbol{b}_{Yj}(t)^\top = (\hat{\phi}_1^E(t), \ldots, \hat{\phi}_{K_{Yj}}^E(t))$ in practice, truncating the infinite sum at a finite number K_{Yj}. FPCs provide interpretable information on the main modes of variation in the data, as the eigenvalues represent the amount of variation explained by each component and the eigenfunctions represent the shape of this variation. Principal components have the advantage of often yielding a small parsimonious basis due to their optimal approximation property for a given number of basis functions. This is a big advantage in the case of large n, where using a large number of

penalized spline basis functions for each E_i is computationally expensive. Due to the link between quadratic penalties and Gaussian distributional assumptions and the resultant equivalence of our general model term representation (3.2) and (3.3) with GP priors (cf. equation (3.4)), $\theta_{j,il} \sim \mathcal{N}(0, \kappa_l^E)$ independently motivates the choice of $\boldsymbol{P}_{xj} = 0$ and $\boldsymbol{P}_{Yj} = \text{diag}(\widehat{\kappa}_1^E, \ldots, \widehat{\kappa}_{K_{Yj}}^E)^{-1}$ here, with λ_{Yj} fixed to 1. This further reduces computational complexity as there is no need for smoothing parameters estimation. Despite the need to first estimate the FPC decomposition, this leads to a pronounced computational advantage of FPC bases over spline bases in this setting (cf. Cederbaum et al., 2016).

For a simple smooth residual model such as model (1.1), estimation of the eigenfunctions can be based on first estimating the mean structure under a working i.i.d. assumption along t and then smoothing the empirical covariance of the centred process leaving out the diagonal, which is contaminated by the variance of ε_{it} (Staniswalis and Lee, 1998; Yao et al., 2005a). In the models displayed in Figure 2, for example, we used FPC decompositions of the smoothed empirical covariance of pilot estimates of E_i, obtained from fitting the model under a working assumption of i.i.d. errors along t. The FPC basis was then used to model the observed autocorrelation and heterogeneous variance along t with a compact basis representation.

Di et al. (2009), Greven et al. (2010), Shou et al. (2015) and Cederbaum et al. (2016) discuss the estimation of the eigenfunctions and eigenvalues for more complex functional random effects models. For example, with more data, we could add curve-specific functional random intercepts E_{aw} to model (1.4), which would be nested within subject-specific functional random intercepts B_a. More generally, such models can contain several (partially) crossed or nested random intercepts or slopes, for grid or sparse functional data. Zipunnikov et al. (2011, 2014) discuss the extension to image data. The general idea is to use cross-products $Y_i(s)Y_{i'}(t)$ as estimators of $\text{Cov}(Y_i(s), Y_{i'}(t))$ after estimation of the mean structure and centring, and then to decompose this covariance into the additive contributions from the random intercepts and slopes, smooth residuals and additional white noise, using a least squares approach or a corresponding additive bivariate varying coefficient model. Smoothing of covariances and an eigendecomposition on a grid of values in $\mathcal{T}$ then yields estimates of the eigenfunctions for each random process in the model. The smoothing step can be adapted to the smoothness of the data and could in principle also be done using other bases than splines. For at most two nested functional random intercepts and scalar covariates, alternatives exist—some of these for generalized functional responses—that directly estimate the FPCs under orthonormality constraints within one overall model (e.g., James et al., 2000; Van der Linde, 2009; Peng and Paul, 2012; Goldsmith et al., 2015).

For expansion of all model terms $h_j(\boldsymbol{x})$ over t, we here focus on the two approaches using spline bases and FPCs. The reason is that both of these can be cast into a quadratic penalty framework as in equation (3.3), which allows reducing the problem to a known penalized estimation problem for scalar data amenable to inference using, for example, mixed models or component-wise gradient boosting. This in turn enables

the use of a broad spectrum of existing statistical methods for the flexible modelling of scalar data.

For functional covariates, there is a corresponding choice of basis. In the scalar-on-function model (1.2), different basis functions $\Phi_{xj,k}$ can be used in the basis expansion (3.1) of the β-coefficient function. Again, we use either splines with a smoothness penalty or an FPC basis, now computed for the covariate process. If both x and β are represented in an expansion using the eigenfunction basis of x, then the regression problem simplifies to (Müller and Stadtmüller, 2005)

$$h_j(\boldsymbol{x}_i) = \int_{\mathcal{S}} x_i(s)\beta(s)ds = \int_{\mathcal{S}} \left\{ \sum_{k=1}^{\infty} \xi_{ik}\phi_k^X(s) \right\} \left\{ \sum_{e=1}^{\infty} \theta_{j,e}\phi_e^X(s) \right\} ds = \sum_{k=1}^{\infty} \xi_{ik}\theta_{j,k}, \qquad (3.5)$$

due to the orthonormality of the eigenfunctions ϕ_k^X of the covariate process. After truncation at a suitable level K_{xj}, expression (3.5) corresponds simply to a regression onto the (estimated) scores ξ_{ik}, with unknown coefficients $\theta_{j,k}$. Then, $\boldsymbol{b}_{xj}(\boldsymbol{x}_i)^\top = (\widehat{\xi}_{i1}, \ldots, \widehat{\xi}_{iK_{xj}})$ and, for example, $\boldsymbol{P}_{xj} = \mathbf{0}$ (for an alternative penalizing β away from directions with little variation in x, see James and Silverman (2005)). A similar approach could be taken with other basis expansions for x and β, for example, using wavelets (see Meyer et al., 2015). In the function-on-function regression model (1.3) with coefficient $\beta(s,t)$, the coefficients $\theta_{j,k}$ for ξ_{ik} are replaced by $c_{j,k}(t)$, and we can estimate each using a spline or FPC (cf. Yao et al., 2005b) basis expansion $c_{j,k}(t) = \sum_{l=1}^{K_{Yj}} \phi_{Yj,l}(t)\theta_{j,kl}$, changing $\boldsymbol{b}_{Yj}(t)^\top$ from 1 to $(\phi_{Yj,1}(t), \ldots, \phi_{Yj,K_{Yj}}(t))$. For an extension and a comparison of both FPC-based and spline-based scalar-on-function regression when responses and covariates are longitudinally observed, see Gertheiss et al. (2013a).

3.5 Application example

As an example to illustrate the model terms and basis expansions discussed in this section, we consider a longitudinal and generalized extension of model (1.2),

$$g(\mathbb{E}(Y_{aw}|\boldsymbol{X}_{aw} = \boldsymbol{x}_{aw})) = \alpha + B_a + \int_{\mathcal{S}} \tilde{x}_{aw}(s)\beta(s)ds + \gamma(z_{aw}), \qquad (3.6)$$

with Y_{aw} now the PASAT score for MS patient a at visit w, $B_a \sim \mathcal{N}(0, \sigma_b^2)$ a subject-specific random intercept, $\tilde{x}_{aw}(s)$, $s \in \mathcal{S}$ the corresponding mean-centred FA–CCA profile at the wth visit, z_{aw} the time in days since the first visit and g a fixed link function. These model terms were selected by stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013) from a much larger model fitted by component-wise gradient boosting for functional data (cf. Section 4.2). Section 4.3 revisits the example and gives details on the boosting results; this section summarizes the results for the mixed model-based inference approach described in Section 4.1.

PASAT scores range from 0 to 60 and count the number of times subjects correctly add consecutive pairs of numbers as they listen to a series of numbers being read to them. Since difficulty of the addition task may increase with prolonged duration due to fatigue, assuming a conditional binomial distribution with 60 identical, independent trials for the score seems questionable. Instead, we divide the raw scores by 60, treat them as quasi-continuous and use a beta distribution for the 'proportion' of correct responses for our model, with a logit link-function. Both refund's `pffr` for functional response regression and `pfr` (Reiss and Ogden, 2007; Goldsmith et al., 2012) for scalar-on-function regression can model many response distributions outside the exponential family using the implementation of Wood et al. (2016b) in R package mgcv.

We fit the model on a stratified training sample using at least two visits of each subject, for a total of 243 out of 340 available observations. For the linear functional effect $\int_{\mathcal{S}} \tilde{x}_{aw}(s)\beta(s)ds$, we compare a model specification where β is represented in terms of 10 cubic B-spline basis functions with first-order difference penalty with an FPC-based one as in equation (3.5). The first order difference penalty imposes a weakly informative prior that the effect is constant over $\mathcal{S}$, which corresponds to an assumption that only the average deviation of FA–CCA profiles from their sample mean curve affects PASAT scores.

Both AIC-based selection of the number of FPCs on the training set and optimization of prediction performance on the test set yield models with 34 FPCs, although the exact optimal number varies for different splits in training and test datasets. In any case, improvements in predictive accuracy are small for larger FPC bases, while the coefficient function's shape and that of its confidence band change quite substantially. For less than 14 FPCs, the coefficient function is very similar to the spline-based estimate. As is often the case, the discrete regularization parameter for this type of effect, that is, the number of leading FPCs to retain in the model, is difficult to optimize.

Figure 3 displays results for the two fits, which yield very similar predictive accuracy on the validation sample: the mean predictive negative log-likelihood (MSE) is 2.986 (0.0099) for the spline-based fit and 2.988 (0.0100) for the FPC-based fit.

In both models, the random subject effect (left panel) is the biggest contributor to the additive predictor by far, followed by the effect of time since first visit and the effect of FA–CCA. Absolute effect sizes for FA–CCA in the FPC-based model are about half of the estimated effect sizes in the spline-based model. The estimated effect of time since first visit z (middle panel) indicates that PASAT scores tend to increase over time and then level off, with some evidence for a subsequent decrease towards the end of the follow-up period. A possible interpretation is that the learning effect for the task is stronger than disease-related deleterious effects on cognitive performance over most of the follow-up, which is rather short compared to the typical speed of MS progression. Due to the low average number of replicates per subject and the large between-subject variance of PASAT scores, these models are very parameter-intensive: Both model fits and a non-linear variant have ≈ 97 effective degrees of freedom (edf) based on just 243 observations and 120 (linear spline-based), 144 (FPC-based) or 134 (non-linear spline-based) coefficients in their $\boldsymbol{\theta}_3$, respectively. The uncertainty about

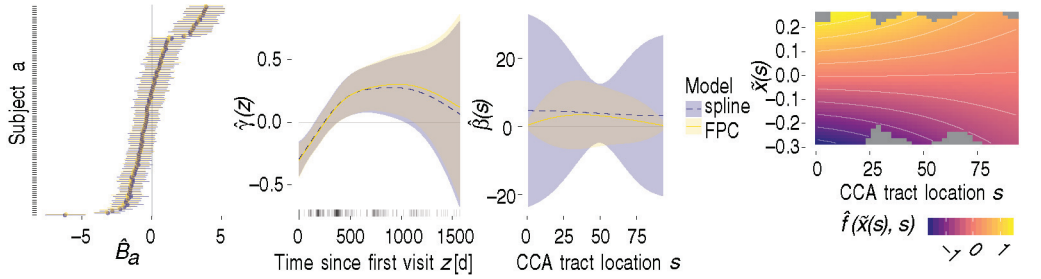

Figure 3 Model (3.6) and a non-linear extension, from left: Predicted subject random effects $\hat{B}_a$ (sorted by value); estimated effects $\hat{\gamma}(z)$ of time since first visit z (in days) along with a rug plot of the observed z_{aw} at the bottom; estimated coefficient functions $\hat{\beta}(s)$ for the linear association with centred FA–CCA curves $\bar{x}(s)$, $s \in \mathcal{S}$; estimated non-linear response surface $\hat{f}(\bar{x}(s), s)$. Dark grey areas in rightmost plot are outside of data support. Intervals are ±2 standard errors, left three panels show results for both FPC- (in gold) and spline-based (in blue) functional linear models.

the effect of FA–CCA is too large to permit reliable substantial interpretation. Taking the point estimates at face value, we would conclude here that the association between FA–CCA and PASAT scores is not strongly localized, since $\hat{\beta}$ is rather constant, and that larger FA (i.e., more unidirectional liquid diffusion and therefore better neuronal health) all along the CCA is associated with higher PASAT scores indicating higher cognitive ability, since $\hat{\beta}(s) > 0$ for all s.

Note that despite the fairly similar point estimates $\hat{\beta}$, the point-wise confidence intervals (CIs) are vastly different for FPC- and spline-based effects. The CI for the FPC-based fit is much narrower since it implicitly conditions on the empirical FPCs. The uncertainties about the estimated FPC representation and its truncation parameter K_{xj} are not included (see Goldsmith et al. (2013) for more details and resampling-based remedies in a functional response context). FPC-based CIs are also shaped very differently than the spline-based ones as these two basis representations imply vastly different (prior) assumptions for the shape of β on different functional spaces spanned by these basis functions, which strongly affects the (posterior) covariance of the estimated coefficient functions.

Non-linear effects $\int f(\tilde{x}(s), s)ds$ can be represented via marginal basis vectors $\boldsymbol{b}_{xj}(\boldsymbol{x})^\top = \left(\sum_{r=1}^{R} \Delta(s_r) \left(\boldsymbol{\Phi}_{sj}(s_r) \otimes \boldsymbol{\Phi}_{xj}(\tilde{x}(s_r))\right)\right)$, which are constructed by numerically integrating $K_{xj} = K_{sj}K_{\tilde{x}j}$ tensor product basis functions of a marginal basis $\boldsymbol{\Phi}_{sj} = (\Phi_{sj,k_s})_{k_s=1,\ldots,K_{sj}}$ over $\mathcal{S}$ and a marginal basis $\boldsymbol{\Phi}_{xj} = (\Phi_{xj,k_x})_{k_x=1,\ldots,K_{\tilde{x}j}}$ over the range of $\tilde{x}(s)$ (McLean et al., 2014). The estimated $\hat{f}$ shown in Figure 3 is based on $K_{\tilde{x}j} = K_{sj} = 5$ marginal cubic B-spline basis functions with second-order difference penalty in x-direction penalizing deviations from a linear effect and a first-order difference penalty over s penalizing deviations from a constant effect. In the small data example discussed here, the uncertainty associated with this complex effect is of the same magnitude as the absolute values of $\hat{f}(\tilde{x}(s), s)$. The predictive performance of this model is equivalent to that of the simpler linear models described above. The

point estimate $\hat{f}$, shown in the right most panel of Figure 3, is roughly linear in $\tilde{x}(s)$ with a positive slope like $\hat{\beta}$ over the entirety of $\mathcal{S}$, albeit with a much smaller slope for $s > 80$. In this case, the data do not seem to strongly indicate a non-linear effect of $\tilde{x}$ and the effect is reduced to the simpler linear case by the penalization. The contributions of $\int \hat{f}(\tilde{x}(s), s)ds$ to the additive predictor are practically identical to those of $\int \tilde{x}(s)\hat{\beta}(s)ds$ in the spline-based linear model.

4 Estimation

After expanding all model terms in penalized basis expansions, the resulting penalized regression model can be estimated using different approaches. We introduce in particular two estimation procedures based on mixed models in Section 4.1 and boosting in Section 4.2, and discuss alternatives in Section 4.4. The key idea is that our modelling approach effectively models the single observations within curves and shifts the functional structure to the additive predictor (2.1), including smooth residual terms to model auto-correlation and heterogeneous variance along t where necessary. This means that the resulting penalized regression is equivalent to a regression problem for 'scalar data', so that we can directly build on the advances in flexible models for such data that have been achieved over the last years and decades and will also be able to utilize future developments.

The two alternatives for estimation we discuss in the following can be seen as complementary, each with its own advantages and disadvantages. The mixed model-based approach (Scheipl et al., 2015, 2016 building on Wood et al., 2016b; see also, e.g., Goldsmith et al., 2012; Ivanescu et al., 2015) can be used when $\boldsymbol{\xi} = g \circ \mathbb{E}$ for some link function g, and with the assumption that conditional on the additive predictor $h(\boldsymbol{x})$, the observed response values (independently within and across functions) come from an exponential family distribution or one of several others like beta or scaled and shifted t-distributions. This approach works well with a moderate number of covariates and has the advantage of providing likelihood-based inference. Extensions of this approach to GAMLSS models have been discussed by Brockhaus et al. (2016a) for signal regression and a few response distributions.

The component-wise gradient boosting approach (Brockhaus et al., 2015b, 2016b building on Hothorn et al., 2014) can handle very general loss functions and thus allows for, for example, mean, median or quantile regression, as well as robust regression or even generalized additive models for location, scale and shape (GAMLSS; Rigby and Stasinopoulos, 2005) modelling several parameters of the conditional response distribution simultaneously (Brockhaus et al., 2015a). The iterative, component-wise estimation algorithm means that many covariates can be handled, even more than observed curves, and that model terms are automatically selected or deselected during estimation. A disadvantage of boosting is that currently no inference for the estimated effects is directly available and resampling methods have to be used for uncertainty quantification.

In cases where both estimation approaches can be applied, they are expected to yield similar results. Brockhaus et al. (2016a) find comparable performance for the two in a simulation-based comparison for a particular GAMLSS signal regression setting, although boosting shows a stronger shrinkage effect in situations with little information content of the data. For example, comparing the mixed model-based estimates for model (3.6) shown in Figure 3 with the corresponding boosting results (cf. the online appendix), we see much stronger regularization in the latter approach for the subject random effects and the effect of the time since first visit but not for the effect of FA–CCA, while the basic structure of the effect estimates is qualitatively similar.

4.1 Mixed model-based inference

If our transformation function can be written as $\boldsymbol{\xi} = g \circ \mathbb{E}$ for a link function g, we can write our model for $\boldsymbol{Y} = (Y_{11}, \ldots, Y_{1D_1}, \ldots, Y_{n1}, \ldots, Y_{nD_n})^\top$ as

$$g(\mathbb{E}(\boldsymbol{Y})) = \mathfrak{X}\boldsymbol{\theta},$$

with the design matrix $\mathfrak{X}$ containing the entries for $(\boldsymbol{b}_{xj}(\boldsymbol{x}_i, t_{id})^\top \odot \boldsymbol{b}_{Yj}(t_{id})^\top)$ ranging over $i = 1, \ldots, n$ and $d = 1, \ldots, D_i$ in rows, and concatenating design matrices column-wise for the partial effects $h_j(\boldsymbol{x})$, $j = 1, \ldots, J$. The vector of unknown coefficients $\boldsymbol{\theta}$ contains blocks $\boldsymbol{\theta}_j$ of coefficients for each j and the link function g is applied entry-wise. Let $\boldsymbol{\lambda}$ be the vector containing all smoothing parameters in $\boldsymbol{P}_j, j = 1, \ldots, J$.

For given smoothing parameters $\boldsymbol{\lambda}$, the coefficients $\boldsymbol{\theta}$ are estimated by maximizing the penalized log-likelihood

$$l(\boldsymbol{\theta}, \boldsymbol{\rho}) - \frac{1}{2}\sum_{j=1}^{J} \boldsymbol{\theta}_j^\top \boldsymbol{P}_j \boldsymbol{\theta}_j,$$

where the log-likelihood $l(\boldsymbol{\theta}, \boldsymbol{\rho})$ is obtained from the assumed conditional density of $\boldsymbol{Y}$ given $\mathfrak{X}$, possibly depending on a vector of nuisance parameters $\boldsymbol{\rho}$, and assuming independence of the Y_{id} within and across functions conditional on the additive predictor. Note that each $\boldsymbol{P}_j$ depends on the respective smoothing parameters λ_{xj} and λ_{Yj}, see equation (3.3).

We follow the approach of Wood (2006a, 2011) and Wood et al. (2016b) for optimization of this penalized log-likelihood and determination of the smoothing parameters (see Scheipl et al., 2015, 2016 for more details). To estimate the smoothing parameters $\boldsymbol{\lambda}$, we use the marginal likelihood with respect to $\boldsymbol{\lambda}$, integrating $\boldsymbol{\theta}$ out of the penalized likelihood based on the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{\theta}$ when interpreting the penalty as a distributional assumption on $\boldsymbol{\theta}$. An estimate for $\boldsymbol{\lambda}$ is then obtained by maximizing a Laplace approximate version of this marginal likelihood (Wood et al., 2016b).

We build on the methods for GAMMS available in the R package mgcv (Wood, 2016b) for our implementation in the `pffr` function of the R package refund. One

advantage of this approach is the availability of CIs and tests using mixed model-based likelihood inference methodology (e.g., Marra and Wood, 2012; Wood, 2013), with close to nominal CI coverages for (generalized) functional responses on simulated data (Ivanescu et al., 2015; Scheipl et al., 2015, 2016). Note that if estimated FPC bases are used, then inference is conditional on this basis and does not include its estimation uncertainty. Cederbaum et al. (2016) found coverage of confidence bands to be close to nominal in simulations for such settings; see Goldsmith et al. (2013) for a resampling-based adjustment in a simpler functional response setting.

4.2 Component-wise gradient boosting

The underlying idea for estimation using boosting is to represent the estimation problem for model (2.1) as a minimization problem of a corresponding loss function that does not necessarily imply a conditional distributional assumption about the responses. Common loss functions include the squared error loss for mean regression ($\boldsymbol{\xi} = \mathbb{E}$), the absolute loss for median regression ($\boldsymbol{\xi} = \text{median}$), the check function for quantile regression ($\boldsymbol{\xi} = q_\tau$ for some τ-quantile) and the negative log-likelihood for responses of the exponential family ($\boldsymbol{\xi} = g \circ \mathbb{E}$ for a link function g). To obtain a suitable loss function for functional responses, we integrate the point-wise (potentially weighted) scalar loss function at each t over $\mathcal{T}$, giving a scalar $L((Y, \boldsymbol{x}), h)$ measuring the loss for response Y and predictor $h(\boldsymbol{x})$. This corresponds to point-wise mean regression, median regression, etc.

The goal then is to minimize the expected loss, the risk, with respect to the predictor h. Component-wise gradient boosting (Bühlmann and Hothorn, 2007; Hastie et al., 2011; Hothorn et al., 2014) can be seen as a gradient descent approach in function space. The empirical risk is iteratively minimized in the direction of the steepest descent (negative gradient) with respect to h. $\hat{h}$ is updated along an estimate of the negative gradient in each step, fitting the negative gradient U_i for all observations $i = 1, \ldots, n$ using so-called base learners, in our context corresponding to J penalized regression models for the partial effects $h_j(\boldsymbol{x}_i)$. Then, the $j^\star$ of the best fitting base learner in this iteration is selected and only the coefficients for this $h_{j^\star}$ are updated. The estimate for h in step m is then given by $\widehat{h}^{[m]} = \widehat{h}^{[m-1]} + \nu \widehat{h}_{j^\star}^{[m]}$, where $\widehat{h}_{j^\star}^{[m]}$ corresponds to the estimate for $h_{j^\star}$ in this step and ν is a step length in $(0, 1)$. The final $\hat{h}^{[m_{\text{stop}}]}$ is a linear combination of base learner fits, reflecting the ensemble nature of the boosting algorithm. Changing from scalar to functional responses means that U_i and $h_j(\boldsymbol{x}_i)$ are now both functions over $\mathcal{T}$. An extension to GAMLSS-type models with multiple additive predictors for modelling more than one feature of the response distribution is described in Mayr et al. (2012) for the scalar case. Brockhaus et al. (2015a, 2016a) extend this approach to functional responses and covariates.

There are several tuning parameters for this algorithm. The smoothing parameters $\boldsymbol{\lambda}$ are chosen and fixed such that the degrees of freedom per iteration are the same for each baselearner. This is important to ensure comparability and unbiased selection of base learners (Kneib et al., 2009; Hofner et al., 2012). The step length ν is usually

fixed to a small number such as 0.1. The number of iterations m_{stop} then controls the model complexity, with large m_{stop} leading to more complex models and lower values (early stopping) corresponding to stronger regularization of estimates. The stopping iteration is chosen by resampling methods such as cross-validation or bootstrapping on the level of curves (or larger independent units such as curves for one subject).

Variable selection results from both early stopping—not all base learners are selected in at least one of the iterations—and additional stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013), which is based on the stability of base learner selection under sub-sampling. The component-wise nature of the algorithm also means that the full model is never fit, only partial models are, including single terms $h_j(\boldsymbol{x})$. This is the reason why this estimation approach can handle more variables than observations.

This approach is implemented in the R package FDboost (Brockhaus and Rügamer, 2016) and builds on model-based boosting as implemented in the R packages mboost (Hothorn et al., 2016) and gamboostLSS (Hofner et al., 2016), exploiting the Kronecker product structure of expansion (3.2) in the case of regular grids and t-constant covariates to increase computational efficiency following Hothorn et al. (2014). For more details, see Brockhaus et al. (2015b).

4.3 Application example

While Section 3.5 describes estimation and inference results for a comparatively simple model estimated in the mixed model framework of Section 4.1, this section illustrates how the broad range of response distributions as well as consistent model selection via stability selection, which are both available for the boosting implementation of Section 4.2 in package FDboost, allow us to explore a large number of possibly quite complex models for the PASAT scores.

Specifically, the maximal model we select from includes a sex effect, a random intercept for the subjects, a non-linear effect of time since first visit in days, a random linear slope for time since first visit for the subjects, linear effects of functional covariates FA–CCA and FA–RCST, linear effects of the derivatives of FA–CCA and FA–RCST as approximated by simply taking the first differences, as well as the interaction effects between sex and the four functional covariates and between sex and time since first visit.

Since it is unclear which distribution is the most appropriate for modelling the PASAT scores, we compare model fit and selected terms for binomial, beta and beta–binomial models, as well as a (distribution-free) median regression model. For the beta and beta–binomial models, we investigate models that model only the expected value as well as those that also feature an additional additive predictor for the dispersion parameter.

Optimal stopping iterations for each model are determined by evaluating the out-of-bag risk for 100 bootstrap samples of the data. Since the risk function that is minimized for these models is simply the negative log–likelihood of the respective model, we can determine the most appropriate distributional assumption

by comparing average predictive risk (i.e., mean negative log likelihood over the 100 bootstrap samples) at the optimal stopping iteration. We also use stability selection with a cut-off at probability of inclusion of 80% and a maximal per-family error rate of 1 for variable and term selection.

Table 1 Comparison of boosting models. From top to bottom: binomial model, beta and beta–binomial models with constant variance, beta and beta–binomial models with modelled variance, median regression. Columns show model name, mean predictive risk and model terms selected by stability selection for each additive predictor. 'Visit time' is the time since first visit z. The code supplement includes exemplary plots of the estimated effects for the $Be(\mu, \sigma = \text{const})$ model.

Model	Risk	Selected
$B(n = 60, p = \mu)$	3.92	FA-CCA
$Be(\mu, \sigma = \text{const})$	3.05	FA-CCA
$BB(\mu, \sigma = \text{const}, n = 60)$	3.33	FA-CCA, visit time
$Be(\mu, \sigma)$	3.01	μ: FA-CCA; σ: $\frac{d}{ds}$ FA-CCA
$BB(\mu, \sigma, n = 60)$	3.32	μ: FA-CCA, visit time; σ: sex, visit time
q_{50}	3.70	FA-CCA

Table 1 shows results for the six models. The beta regression model with constant dispersion parameter achieves an average risk around 3.05, while a beta regression model with a modelled dispersion parameter achieves around 3.01. Models based on other distributions perform worse. Note that median regression is analogous to mean regression for a conditionally Laplace distributed response.

For most models, just the linear effect of FA–CCA on the conditional mean is selected. Only the beta–binomial models additionally include the non-linear effect of time since first visit. Stability selection for the beta regression with modelled dispersion selects a linear effect of the derivative of FA–CCA for modelling the dispersion, while the beta-binomial model with modelled dispersion selects sex and time since first visit. The mixed model-based beta regression model described in Section 3.5 additionally includes an effect of time since first visit, since the effect is quite close to the threshold of selection for the boosted beta model. It also additionally includes subject-specific random intercepts that are not stability selected but serve to model the dependence structure of these longitudinal data.

4.4 Alternatives

Model (2.1), with the penalized basis representations discussed in Section 3, can also be estimated by other methods such as a Bayesian approach. Bayesian implementations exist for certain special cases, for example, Goldsmith et al. (2011, 2015). Much more generally, all exponential family models available in `pffr` can be automatically translated into JAGS (Plummer, 2016) code using mgcv's `jagam` function (Wood, 2016a) for automated, tuning-free, fully Bayesian MCMC inference.

A very general Bayesian alternative is Morris and Carroll (2006), Morris et al. (2011) and further work by this group, partially implemented in the WFMM (Herrick, 2015) software suite. They take a different approach for estimation from the one focused on here, first transforming the response curves using a (usually wavelet)

basis transformation and subsequently building a model in the basis coefficient space using variable selection priors to allow zeros for wavelet coefficients, before transforming results back into function space. Advantages of this approach are the ability of wavelets to handle spiky data, the availability of extensions to image data and the computational efficiency of parsimonious but lossy transformations for very high-dimensional data sets. Disadvantages are that grids need to be identical for all functional responses, some model terms such as historical functional effects cannot be estimated and the publicly available implementation is restricted to mean regression for Gaussian functional responses.

5 Challenges and technical points

5.1 Necessary constraints in models with functional responses

In regression with functional responses and/or covariates, identifiability has to be carefully considered. The first point concerns functional response models, including at least two $h_j(\boldsymbol{x})$ varying over t. This is most easily seen when the model includes a smooth intercept. As

$$\alpha(t) + h_j(\boldsymbol{x})(t) = [\alpha(t) + \bar{h}_j(t)] + [h_j(\boldsymbol{x})(t) - \bar{h}_j(t)] =: \widetilde{\alpha}(t) + \widetilde{h}_j(\boldsymbol{x})(t),$$

with $\bar{h}_j(t) = \frac{1}{n}\sum_{i=1}^{n} h_j(\boldsymbol{x}_i)(t)$, a constraint on $h_j(\boldsymbol{x})$ is needed to ensure identifiability, where a straight forward interpretation is achieved with the constraint $\bar{h}_j(t) = 0$ for all t. Such a constraint can be incorporated by appropriately modifying the design matrix (cf. Wood (2006a); see Brockhaus et al. (2015b) on how to preserve the Kronecker structure for tensor product bases as in expansion (3.2)).

5.2 Identifiability for functional covariates

The second point is more serious, as it is less easily remedied. It concerns the estimation of effects of functional covariates (see Scheipl and Greven (2016) for a more detailed discussion and further references). Consider for simplicity a linear functional effect as in model (1.2). It is often the case in practice that the functional covariate can be well approximated by the first M components of the Karhunen–Loève expansion,

$$x_i(s) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k^X(s) \approx \sum_{k=1}^{M} \xi_{ik}\phi_k^X(s),$$

as in any case at most $\min(n, R)$ eigenfunctions ϕ_k^X with non-zero eigenvalues can be estimated from the data.

If we use an FPC basis for β with $K_{xj} \leq M$, then the coefficients $\theta_{j,k}$ in term (3.5) are identifiable. It is important to note, however, that this is contingent on the assumption that β lies in the function space spanned by the first K_{xj} eigenfunctions of x. If the

eigenfunctions are non-smooth, as higher-order eigenfunctions often are, then this can lead to non-smooth β-function estimates. The estimated shape of β is also typically highly dependent on the chosen K_{xj} (see Goldsmith et al. (2012), and the discussion of the FPC-based effect of FA–CCA in Section 3.5).

If alternatively β is assumed to be smooth in s, then one would usually use a basis of spline functions $\Phi_{xj,k}$, with the basis vector $\boldsymbol{b}_{xj}(\boldsymbol{x})$ containing the terms $\sum_{r=1}^{R} \Delta(s_r)x(s_r)\Phi_{xj,k}(s_r)$, $k = 1, \ldots, K_{xj}$. Let $\boldsymbol{\phi}^X = (\phi_k^X(s_r))_{k=1,\ldots,M;r=1,\ldots,R}$, $\boldsymbol{\Delta} = \text{diag}(\Delta(s_1), \ldots, \Delta(s_R))$ and $\boldsymbol{\Phi}_{xj} = (\Phi_{xj,k}(s_r))_{r=1,\ldots,R;k=1,\ldots,K_{xj}}$. If $M < K_{xj}$ or $\text{rank}(\boldsymbol{\phi}^X\boldsymbol{\Delta}\boldsymbol{\Phi}_{xj}) < K_{xj}$, then the functional covariate does not contain sufficient information to uniquely determine the K_{xj} spline basis coefficients and the resulting design matrix will be rank-deficient. (Near-deficient design matrices will result in large condition numbers and estimates with large variability.) In that case, the penalized (likelihood) criterion is minimized by the smoothest solution among all possible solutions with equally good fit to the data. This smoothest solution will be unique as long as the null space of the penalty does not overlap the null space of the design matrix.

This leads to several practical recommendations. The first is to avoid rank-reducing pre-processing of functional covariates such as pre-smoothing or curve-wise centering, where possible. The second is to compute diagnostic measures to check for any problems in practice; such measures are implemented in the refund and FDboost packages. And the third is to keep the null space of the penalty small by using first-order differences or derivatives (penalizing deviations from a constant coefficient function), rather than higher-order differences or derivatives. An alternative is to apply constraints on the coefficient function that force its components in the overlap of the design matrix null space and the penalty null space, where there is no information on the coefficient shape available from the data or the prior/penalty, to zero. As this corresponds to an implicit assumption that β does not have a non-zero component in this overlap, corresponding warnings are issued in our implementation if such a constraint is used.

Functional covariates in our application example are of high rank, and the checks for identifiability that `pffr` automatically performs indicate no identifiability issues here. However, the example in Section 3.5 also demonstrates that even for such non-pathological cases, assumptions on the shape of the functional coefficients expressed through their penalized basis representations can strongly affect estimated coefficient shapes.

5.3 Computational complexity

There are two important ways in which computational efficiency of the model fitting can be increased. First, the choice of the basis can be important for the computational complexity. We discussed in Section 3.4 how FPCs are a useful basis in cases such as functional random effects or smooth residuals. For those terms, fitting a smooth curve for each factor level or observation means that the number of basis functions

gets multiplied by the number of factor levels for the random effect or by n for the smooth residual. Not only is the number of basis functions reduced for the more parsimonious FPC basis, but also the smoothing parameters are not estimated within the large overall model. In particular, this can help to speed up mixed model-based inference.

The second important approach can be employed if all response curves are observed on the same grid, which need not be equidistant. Some missing observations within curves are also allowed and taken into account with zero weighting. The second requirement is that the covariate values do not change with t, that there are in particular no concurrent or historical functional effects in the model. Then, the tensor product basis representation (3.2) rather than the more general row tensor product basis can be used. This leads to Kronecker products in the design matrix and together with the Kronecker sum penalty (3.3) results in a special case of the generalized linear array model introduced by Currie et al. (2006). Their approach defines very efficient array-based operations on the much smaller marginal design matrices to compute linear functions and quadratic forms of the overall design matrix, which is never computed explicitly, and can thus significantly decrease both computation time and memory footprint. As our implementations rely on existing software implementations of flexible models for scalar data, these array model methods are currently implemented for the boosting approach in FDboost based on the mboost package, but not in the `pffr` function in the refund package based on mgcv. While for smaller datasets, `pffr` can be much faster than FDboost due to FDboost's need for repeated fits on resampled data sets to select the stopping iteration, for larger data sets and models, computation time for FDboost tends to increase much more slowly—especially for the array case (Brockhaus et al., 2015b), but also in the case of many covariate terms due to fitting each base learner separately.

On a modern desktop PC, fitting the functional response model (1.1) (cf. Figure 2) with refund's `pffr` function required about 4 minutes for the Gaussian model, about 7 minutes for the beta model with logit link and about 30 minutes for the Gaussian model with heteroskedastic residuals. Fitting a simpler model without smooth residuals took about 4 seconds for the Gaussian and beta models and 28 seconds for the heteroskedastic Gaussian model. For the scalar response model (1.2) (cf. Figure 3, Table 1), single fits of FDboost without early stopping took between 5 and 10 seconds. We used 100 bootstrap replicates of these full fits to determine suitable stopping iterations m_{stop} and another 100 replicates with early stopping for stability selection. These replicates are performed in parallel, so wall-clock computation times depend largely on the number of available processes for parallelization. The scalar response models shown in Figure 3 required about 1.5 seconds to fit with refund's `pfr` function. These examples illustrate that as both refund and FDboost use iterative algorithms, computation times depend to a large part on the required number of iterations (as well as on the available hardware, of course) and no simple relationship between computation time and data size or model complexity can thus be given.

Recent algorithmic advances for fitting 'giga-scale' additive models on the order of 10^8 data points with 10^4 parameters with the `discrete` option of function `bam` in

mgcv, based on discretizing and compressed storage of continuous predictors (Wood et al., 2016a), are also available via `pffr`.

6 Discussion and outlook

This article discusses a very general framework for regression with functional responses and/or covariates. We hope to have conveyed two key points: First, that many of the particular functional regression models discussed in the literature (cf. Morris, 2015) can be viewed as special cases of this general model class. Second, that if we directly model the observed data points within each function, then we can deploy almost all of the theoretical and computational advances in flexible regression models for scalar data that have been achieved in the last decades for the functional regression setting as well. We believe that both of these points are important for functional regression. The second is essential in building a toolbox for functional regression that is similarly flexible as the toolbox of models available for scalar data without redoing much of that work. And the first, because it allows a unified discussion and implementation of functional regression models. Thus, changing the focus from a scalar-on-function regression to a function-on-scalar regression for a given dataset becomes only a small change in the model call for the data analyst. Even more importantly, as soon as a new feature or model class is developed and implemented for scalar data, it immediately can be made available for all functional data regression models that fit into our framework as well. This strategy has worked well for both boosting-based estimation of GAMLSS models (Hofner et al., 2016) and Laplace approximate marginal likelihood inference for non-exponential family responses (Wood et al., 2016b) in the recent past.

In our formulation, the discussed flexible regression models for functional data are essentially models for the scalar observed points within each function, with the functional data structure shifted to the smooth additive predictor. Thus, asymptotic results on consistency of estimators in such scalar additive models should be applicable to our functional setting as well (see, e.g., Wood et al., 2016b for the type of complex models we consider and mixed model-based inference and Hothorn et al., 2014 for a related result in a boosting context). Simulations (e.g., Brockhaus et al., 2015b; Scheipl et al., 2015, 2016) back up consistency, as well as appropriate coverage of confidence bands relying on asymptotic normality in the mixed model-based case. Nevertheless, this is a point which would deserve closer investigation, particularly regarding regression with functional covariates and with FPC bases as well as regarding different asymptotic regimes of fixed or growing grid sizes D_i with n.

The model class and estimation approaches we discussed are, of course, no panacea for all possible regression settings with functional data. One obvious point here is that while we assume smoothness of underlying curves and effects throughout and our choice of bases and penalties is guided by this assumption, other bases and penalties will be more suitable to other kinds of functional data, for example, spiky data.

Other types of penalties, for example, adaptive penalties, might also be better suited in some situations such as when smoothness is varying along $\mathcal{T}$. To some extent, such adaptive smoothing is available in `pffr` via mgcv's adaptive smoothers. While linear and smooth effects can be easily represented within the discussed framework, more complex relationships require additional work. Principal coordinates (Reiss et al., 2015) have recently been proposed as one way to extend the framework to allow more general non-linear features of functional covariates to influence a response (cf., e.g., Ferraty and Vieu, 2006 for alternatives using a different non-parametric framework). Self-interactions of functions can also be of interest and while Fuchs et al. (2015) could be used for a linear self-interaction model, more complex potentially non-linear relationships would require additional development. We also believe that while more flexible models for functional covariates are of interest and could use further development, both interpretability and identifiability need to be carefully considered and kept in mind, as these are more challenging in our view even for the simple linear regression case than is often appreciated. For functional historical models, estimating the integration limits $[\ell(t), u(t)]$ from the data would often be of interest. Identifiability has to be carefully considered in this setting, and approaches for estimating one endpoint have been proposed in Hall and Hooker (2016); see also Obermeier et al. (2015).

One topic we have not touched on here but that is of large practical importance is that of the registration of curves. Functional regression models assume that the meaning of a point in $\mathcal{T}$ is the same across different functions observed over that interval. However, it is often the case in practical applications that curves need to be 'registered', that is, there is not only variation in 'Y-direction' (amplitude), but also in 't-direction' ('phase'). A classic example is growth data (e.g., Ramsay and Silverman, 2005, Chapter 1), where growth spurts of children not only occur with different intensities but also at different time-points, and the most meaningful comparisons between curves will require aligning the growth spurts in time before further analyses. In fact, the shifted peaks of the smooth residual curves for model (1.1) shown in Figure 2 (second column from left) around the locations of the two global peaks at ca. $t = 10$ and $t = 90$ may indicate less than perfect alignment of the CCA tracts in our data example. While many methods have been developed for curve registration (see Marron et al., 2014, for a recent overview), there is still ample room for methods development jointly looking at registration and flexible regression along the lines of Gervini (2015) or Hadjipantelis et al. (2015).

Finally, we hope that much of what we have learned about functional regression in the last years will lead to fruitful cross-fertilization with other areas of what is sometimes called 'object (oriented) data analysis' (cf. Marron and Alonso, 2014). As the data that is being collected become more and more complex, regression models are required for more general objects, for example, trajectories in 2D or 3D space, images (e.g., Goldsmith et al., 2014) or functions on manifolds (e.g., Ettinger et al., 2016), shapes (e.g., Dryden, 2014) or even objects such as trees (e.g., Wang and Marron, 2007). For all of these, achieving flexible regression models with such responses and/or covariates, random effects, etc. is an active and exciting field of research.

Acknowledgements

This discussion article is based on joint work with Sarah Brockhaus, Brian Caffo, Jona Cederbaum, Ciprian Crainiceanu, Karen Fuchs, Andreas Fuest, Jan Gertheiss, Jeff Goldsmith, Giles Hooker, Torsten Hothorn, Andrada Ivanescu, Fritz Leisch, Andreas Mayr, Mathew McLean, Michael Melcher, Daniel Reich, David Rügamer, David Ruppert, Ana-Maria Staicu, Haochang Shou and Vadim Zipunnikov. We would like to thank all of our co-authors for the fruitful collaboration. We are grateful to the editors Brian Marx, Arnost Komarek and Jeff Simonoff, who initially suggested to write this article, and to Sarah Brockhaus, Jan Gertheiss and Jeff Goldsmith, who read a draft of the article and made very helpful and constructive comments that led to improvements in the manuscript. Financial support was provided by the German Research Foundation (DFG) through Emmy Noether grant GR 3793/1-1. The MRI/DTI data were collected at Johns Hopkins University and the Kennedy-Krieger Institute and we would like to thank Daniel Reich and colleagues for making it publicly available.

References

Brockhaus S, Fuest A, Mayr A and Greven S (2015a) Functional regression models for location, scale and shape with application to stock returns. In H Friedl and H Wagner, eds. *Proceedings of the 30th International Workshop on Statistical Modelling*, vol. 1, Linz, July 6–10, 2015, pages 117–22.

Brockhaus S, Fuest A, Mayr A and Greven S (2016a) Signal regression models for location, scale and shape with an application to stock returns. arXiv preprint arXiv:1605.04281.

Brockhaus S, Melcher M, Leisch F and Greven S (2016b) Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing*. Available at http://link.springer.com/article/10.1007/s11222-016-9662-1. (accessed on 17 November 2016).

Brockhaus S and Rügamer D (2016) *FDboost: Boosting functional regression models*. R package version 0.2–0. Available at http://CRAN.R-project.org/package=FDboost. (accessed on 17 November 2016).

Brockhaus S, Scheipl F, Hothorn T and Greven S (2015b) The functional linear array model. *Statistical Modelling*, **15**, 279–300.

Bühlmann P and Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**, 477–505.

Cardot H, Ferraty F and Sarda P (1999) Functional linear model. *Statistics & Probability Letters*, **45**, 11–22.

Cederbaum J, Pouplier M, Hoole P and Greven S (2016) Functional linear mixed models for irregularly or sparsely sampled data. *Statistical Modelling*, **16**, 67–88.

Currie ID, Durban M and Eilers PH (2006) Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 259–80.

De Boor C (1978) *A Practical Guide to Splines*. New York: Springer.

Di C-Z, Crainiceanu CM, Caffo BS and Punjabi NM (2009) Multilevel functional principal component analysis. *The Annals of Applied Statistics*, **3**, 458–88.

Dryden IL (2014) Shape and object data analysis. *Biometrical Journal*, **56**, 758–60.

Eilers P and Marx B (1996) Flexible smoothing with B-splines and penalties. *Statistical Sciences*, **11**, 89–121.

Eilers PH and Marx BD (2003) Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems*, **66**, 159–74.

Eilers PHC and Marx BD (2002) Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, **11**, 758–83.

Ettinger B, Perotto S and Sangalli LM (2016) Spatial regression models over two-dimensional manifolds. *Biometrika*, **103**, 71–88.

Febrero-Bande M and Oviedo de la Fuente M (2012) Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, **51**, 1–28. Available at http://www.jstatsoft.org/v51/i04/(accessed on 17 November 2016).

Fenske N, Kneib T and Hothorn T (2011) Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association*, **106**, 494–510.

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

Fuchs K, Scheipl F and Greven S (2015) Penalized scalar-on-functions regression with interaction term. *Computational Statistics & Data Analysis*, **81**, 38–51.

Gasparrini A, Armstrong B and Kenward MG (2010) Distributed lag non-linear models. *Statistics in Medicine*, **29**, 2224–34.

Gertheiss J, Goldsmith J, Crainiceanu C and Greven S (2013a) Longitudinal scalar-on-functions regression with application to tractography data. *Biostatistics*, **14**, 447–61.

Gertheiss J, Maity A and Staicu A-M (2013b) Variable selection in generalized functional linear models. *Stat*, **2**, 86–101.

Gervini D (2015) Warped functional regression. *Biometrika*, **102**, 1–14.

Goldsmith J, Bobb J, Crainiceanu CM, Caffo B and Reich D (2012) Penalized functional regression. *Journal of Computational and Graphical Statistics*, **20**, 830–51.

Goldsmith J, Greven S and Crainiceanu CM (2013) Corrected confidence bands for functional data using principal components. *Biometrics*, **69**, 41–51.

Goldsmith J, Huang L and Crainiceanu CM (2014) Smooth scalar-on-image regression via spatial Bayesian variable selection. *Journal of Computational and Graphical Statistics*, **23**, 46–64.

Goldsmith J, Wand MP and Crainiceanu CM (2011) Functional regression via variational Bayes. *Electronic Journal of Statistics*, **5**, 572–602.

Goldsmith J, Zipunnikov V and Schrack J (2015) Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, **71**, 344–53.

Greven S, Crainiceanu CM, Caffo BS and Reich D (2010) Longitudinal functional principal component analysis. *Electronic Journal of Statistics*, **4**, 1022–54.

Greven S, Crainiceanu CM, 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**, 870–91.

Hadjipantelis PZ, Aston JA, Müller H-G and Evans JP (2015) Unifying amplitude and phase analysis: A compositional data approach to functional multivariate mixed-effects modeling of mandarin Chinese. *Journal of the American Statistical Association*, **110**, 545–59.

Hall P and Hooker G (2016) Truncated linear models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78**, 637–53. Available at http://onlinelibrary.wiley.com/doi/10.1111/rssb.12125/abstract

Hastie T and Mallows R (1993) Discussion of 'A statistical view of some chemometrics regression tools,' by I. E. Frank and J. H. Friedman. *Technometrics*, **35**, 140–43.

Hastie T and Tibshirani R (1993) Varying-coefficient models. *Journal of the Royal Statistical Society. Series B*, **55**, 757–96.

Hastie TJ, Tibshirani RJ and Friedman JH (2011) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Herrick R (2015) *WFMM*. The University of Texas M.D. Anderson Cancer Center, version 3.0 edition. Available at https://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=70 (accessed on 16 December 2016)

Hofner B, Hothorn T, Kneib T and Schmid M (2012) A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*, **20**, 956–71.

Hofner B, Mayr A, Fenske N and Schmid M (2016) *gamboostLSS: Boosting Methods for GAMLSS Models*. R package version 1.3-0. Available at http://CRAN.R-project.org/package=gamboostLSS (accessed on 29 October 2016).

Horváth L and Kokoszka P (2012) *Inference for Functional Data with Applications*. New York: Springer Science & Business Media.

Hothorn T, Buehlmann P, Kneib T, Schmid M and Hofner B (2016) *mboost: Model-Based Boosting*. R package version 2.6-0. Available at http://CRAN.R-project.org/package=mboost(accessed on 22 November 2016).

Hothorn T, Kneib T and Bühlmann P (2014) Conditional transformation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**, 3–27.

Huang L, Scheipl F, Goldsmith J, Gellar J, Harezlak J, McLean MW, Swihart B, Xiao L, Crainiceanu C and Reiss P (2016) *refund: Regression with Functional Data*. R package version 0.1–15. Available at https://CRAN.R-project.org/package=refund. (accessed on 22 November 2016).

Ivanescu AE, Staicu A-M, Scheipl F and Greven S (2015) Penalized function-on-function regression. *Computational Statistics*, **30**, 539–68.

James GM, Hastie TJ and Sugar CA (2000) Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.

James GM and Silverman BW (2005) Functional adaptive model estimation. *Journal of the American Statistical Association*, **100**, 565–76.

Karhunen K (1947) *Über Lineare Methoden in der Wahrscheinlichkeitsrechnung*. *Annales Academiae Scientiarum Fennicae*, **37**, 1–79.

Kneib T, Hothorn T and Tutz G (2009) Variable selection and model choice in geoadditive regression models. *Biometrics*, **65**, 626–34.

Koenker R (2005) *Quantile Regression*. Cambridge: Cambridge University Press.

Lang S and Brezger A (2004) Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Loève M (1945) Fonctions aléatoires de second ordre. *Comptes Rendus Académie des Sciences*, **220**, 469.

Malfait N and Ramsay JO (2003) The historical functional linear model. *Canadian Journal of Statistics*, **31**, 115–28.

Marra G and Wood SN (2012) Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, **39**, 53–74.

Marron JS and Alonso AM (2014) Overview of object oriented data analysis. *Biometrical Journal*, **56**, 732–53.

Marron JS, Ramsay JO, Sangalli LM and Srivastava A (2014) Statistics of time warpings and phase variations. *Electronic Journal of Statistics*, **8**, 1697–702.

Marx BD and Eilers PH (1999) Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics*, **41**, 1–13.

Marx BD and Eilers PH (2005) Multidimensional penalized signal regression. *Technometrics*, **47**, 13–22.

Mayr A, Fenske N, Hofner B, Kneib T and Schmid M (2012) Generalized additive models for location, scale and shape for high dimensional data—A flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**, 403–27.

McLean MW, Hooker G and Ruppert D (2015) Restricted likelihood ratio tests for linearity in scalar-on-function regression. *Statistics and Computing*, **25**, 997–1008.

McLean MW, Hooker G, Staicu A-M, Scheipl F and Ruppert D (2014) Functional

generalized additive models. *Journal of Computational and Graphical Statistics*, **23**, 249–69.

Meinshausen N and Bühlmann P (2010) Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 417–73.

Mercer J (1909) Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A.*, **209**, 415–46.

Meyer MJ, Coull BA, Versace F, Cinciripini P and Morris JS (2015) Bayesian function-on-function regression for multilevel functional data. *Biometrics*, **71**, 563–74.

Morris JS (2015) Functional regression. *Annual Review of Statistics and its Applications*, **2**, 321–59.

Morris JS, Arroyo C, Coull BA, Ryan LM, Herrick R and Gortmaker SL (2006) Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: A case study. *Journal of the American Statistical Association*, **101**, 1352–64.

Morris JS, Baladandayuthapani V, Herrick RC, Sanna P and Gutstein H (2011) Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *The Annals of Applied Statistics*, **5**, 894–923.

Morris JS and Carroll RJ (2006) Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, **68**, 179–99.

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

Obermeier V, Scheipl F, Heumann C, Wassermann J and Küchenhoff H (2015) Flexible distributed lags for modelling earthquake data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **64**, 395–412.

O'Sullivan F (1986) A statistical perspective on ill-posed inverse problems. *Statistical Science*, **1**, 502–18.

Peng J and Paul D (2012) A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics*, **18**, 995–1015.

Plummer M (2016) *rjags: Bayesian Graphical Models Using MCMC*. R package version 4-6. Available at https://CRAN.R-project.org/package=rjags. (accessed on 23 November 2016).

Ramsay JO and Dalzell C (1991) Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, **33**, 539–72.

Ramsay J, Graves S and Hooker G (2009) *Functional data analysis with R and MATLAB*. New York: Springer.

Ramsay JO and Silverman BW (2005) *Functional Data Analysis*. New York: Springer.

Ramsay JO, Wickham H, Graves S and Hooker G (2014) *fda: Functional Data Analysis*. R package version 2.4-4. Available at http://CRAN.R-project.org/package=fda. (accessed on 23 November 2016).

Reimherr M and Nicolae D (2016) Estimating variance components in functional linear models with applications to genetic heritability. *Journal of the American Statistical Association*, **111**, 407–22.

Reiss PT and Ogden RT (2007) Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, **102**, 984–96.

Reiss PT, Goldsmith J, Shang HL and Ogden RT (2016) Methods for scalar-on-function regression. *International Statistical Review*. Available at http://onlinelibrary.wiley.com/doi/10.1111/insr.12163/full

Reiss PT, Huang L and Mennes M (2010) Fast function-on-scalar regression with penalized basis expansions. *The International Journal of Biostatistics*, **6**, 1557–4679.

Reiss PT, Miller DL, Wu P-S and Hua W-Y (2015) Penalized nonparametric scalar-on-function regression via principal coordinates. Technical Report. Available at http://works.bepress.com/phil_reiss/42/ (accessed on 23 November 2016).

Reiss PT and Ogden RT (2007) Functional principal component regression and functional partial least squares. Journal of the *American Statistical Association*, **102**, 984–96.

Rigby RA and Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507–54.

Ruppert D Wand MP and Carroll RJ (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.

Scheipl F, Gertheiss J and Greven S (2016) Generalized functional additive mixed models. *Electronic Journal of Statistics*, **10**, 1455–92.

Scheipl F and Greven S (2016) Identifiability in penalized function-on-function regression models. *Electronic Journal of Statistics*, **10**, 495–526.

Scheipl F, Greven S and Küchenhoff H (2008) Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, **52**, 3283–99.

Scheipl F, Staicu A-M and Greven S (2015) Functional additive mixed models. *Journal of Computational and Graphical Statistics*, **24**, 477–501.

Schmid M and Hothorn T (2008) Boosting additive models using component-wise P-splines. *Computational Statistics & Data Analysis*, **53**, 298–311.

Shah RD and Samworth RJ (2013) Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 55–80.

Shi JQ and Cheng Y (2014) *GPFDA: Apply Gaussian Process in Functional data analysis*. R package version 2.2. Available at https://CRAN.R-project.org/package=GPFDA(accessed on 23 November 2016).

Shi JQ and Choi T (2011) *Gaussian Process Regression Analysis for Functional Data*. Boca Raton, FL: CRC Press.

Shi JQ, Wang B, Murray-Smith R and Titterington M (2007) Gaussian process functional regression modeling for batch data. *Biometrics*, **63**, 714–23.

Shou H, Zipunnikov V, Crainiceanu CM and Greven S (2015) Structured functional principal component analysis. *Biometrics*, **71**, 247–57.

Staicu A-M, Li Y, Crainiceanu CM and Ruppert D (2014) Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian Journal of Statistics*, **41**, 932–49.

Staniswalis J and Lee J (1998) Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, **93**, 1403–04.

Swihart BJ, Goldsmith J and Crainiceanu CM (2014) Restricted likelihood ratio tests for functional effects in the functional linear model. *Technometrics*, **56**, 483–93.

Usset J, Staicu A-M and Maity A (2016) Interaction models for functional regression. *Computational Statistics & Data Analysis*, **94**, 317–30.

Van der Linde A (2009) Bayesian functional principal components analysis for binary and count data. *Advances in Statistical Analysis*, **93**, 307–33.

Wand M and Ormerod J (2008) On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, **50**, 179–98.

Wang B and Shi JQ (2014) Generalized Gaussian process regression model for non-Gaussian functional data. *Journal of the American Statistical Association*, **109**, 1123–33.

Wang H and Marron J (2007) Object-oriented data analysis: Sets of trees. *The Annals of Statistics*, **35**, 1849–73.

Wang J-L, Chiou J-M and Mueller H-G (2016) Review of functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257–95.

Wood SN (2006a) *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: CRC Press.

Wood SN (2006b) Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, **62**, 1025–36.

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

Wood SN (2013) A simple test for random effects in regression models. *Biometrika*, **100**, 1005–10.

Wood SN (2016a) Just another Gibbs additive modeller: Interfacing JAGS and mgcv. *arXiv preprint arXiv:1602.02539*. Available at https://arxiv.org/abs/1602.02539(accessed on 23 November 2016).

Wood SN (2016b) *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. R package version 1.8-12. Available at http://CRAN.R-project.org/package=mgcv(accessed on 23 November 2016).

Wood SN, LiZ, Shaddick G and Augustin NH (2016a) Generalized additive models for gigadata: Modelling the UK black smoke network daily data. *Journal of the American Statistical Association*. Available at http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2016.1195744

Wood SN, Pya N and Säfken B (2016b) Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*. Available at http://arxiv.org/abs/1511.03864(accessed on 23 November 2016).

Yao F, Müller H and Wang J (2005a) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–90.

——— (2005b) Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, **33**, 2873–903.

Zipunnikov V, Caffo B, Yousem, DM, Davatzikos, C, Schwartz, BS, and Crainiceanu, C (2011) Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics*, **20**, 852–73.

Zipunnikov V, Greven S, Shou H, Caffo B, Reich DS and Crainiceanu C (2014) Longitudinal high-dimensional principal components analysis with application to diffusion tensor imaging of multiple sclerosis. *The Annals of Applied Statistics*, **8**, 2175–202.