Stat
The ISI's Journal for the Rapid
Dissemination of Statistics Research
(wileyonlinelibrary.com) DOI: 10.100X/sta.0000

Variable Selection in Generalized Functional Linear Models

J. Gertheiss[a]*, A. Maity[b], A.-M. Staicu[b]

Received 00 Month 2013; Accepted 00 Month ????

Modern research data, where a large number of functional predictors is collected on few subjects are becoming increasingly common. In this paper we propose a variable selection technique, when the predictors are functional and the response is scalar. Our approach is based on adopting a generalized functional linear model framework and using a penalized likelihood method that simultaneously controls the sparsity of the model and the smoothness of the corresponding coefficient functions by adequate penalization. The methodology is characterized by high predictive accuracy, and yields interpretable models, while retaining computational efficiency. The proposed method is investigated numerically in finite samples, and applied to a diffusion tensor imaging tractography data set and a chemometric data set.

Keywords: group lasso; multiple functional predictors; penalized estimation; variable selection

1. Introduction

Functional linear regression (Ramsay & Dalzell, 1991) is widely used to explore the relationship between a scalar response and functional predictors. In particular, (generalized) functional linear models are popular in areas including biomedical studies, chemometrics and commerce. In such models, the effect of each functional predictor on the response is modeled through the inner product of the functional covariate and an unknown smooth coefficient function. In this article, we consider the situation where multiple functional predictors are observed, but only a few of these predictors are actually useful in predicting the response. We develop a variable selection procedure to select the important functional predictors and estimate the corresponding coefficient functions simultaneously. Our procedure controls both the sparseness of the regression model and the smoothness of the coefficient functions. Furthermore, the methods can accommodate functional predictors that are measured with error or are observed in a sparse set of points. We investigate the finite sample performance of our procedure via a simulation study and illustrate our method by applying it to a diffusion tensor imaging data set with 30 covariates and a chemometric data set with seven covariates.

Let Y_i denote the scalar response for subject i, and $X_{i1}, \ldots, X_{ip}$ denote the independent realizations of the squared integrable random curves $X_1, \ldots, X_p$, respectively, where $X_j : \mathcal{D}_j \subset \mathbb{R} \to \mathbb{R}$, for $i = 1, \ldots, n$. Without loss of generality,

[a]**Department of Animal Sciences, Georg-August-Universität Göttingen, Germany**
[b]**Department of Statistics, North Carolina State University, USA**
*__Email: jgerthe@uni-goettingen.de__

Prepared using staauth.cls [Version: 2012/05/12 v1.00]

assume that the underlying trajectories X_j's have mean function equal to zero. For simplicity of exposition, consider first the case when X_{ij}'s are observed at a dense grid of time points $\{t_{j1}, \ldots, t_{jN_j}\}$ and do not contain measurement error; this limitation is relaxed in Section 3. In its generality, it is assumed that, given $X_{i1}, \ldots, X_{ip}$, the distribution of the outcome Y_i is in the exponential family with linear predictor η_i and dispersion parameter ϕ, denoted here by $\text{EF}(\eta_i, \phi)$. The linear predictor is set to have the following form

$$\eta_i = \alpha + \sum_{j=1}^{p} \int_{\mathcal{D}_j} X_{ij}(t)\beta_j(t)\, dt, \tag{1}$$

with $E[Y_i|X_{i1}, \ldots, X_{ip}] = \mu_i = h^{-1}(\eta_i)$, where $h(\cdot)$ is a known link function. The coefficient functions $\beta_j(t)$ are assumed as smooth, squared integrable, and represent the main object of interest. Function $\beta_j(t)$ quantifies the effect of the functional predictor $X_{ij}(t)$ on the mean response μ_i. For convenience, throughout the paper the domain of integration $\mathcal{D}_j$ in the integral $\int_{\mathcal{D}_j} X_{ij}(t)\beta_j(t)dt$ if often suppressed. A special case of model (1) corresponds to the identity link function $h(\mu_i) = \mu_i$. The resulting model is known as the functional linear model, and can be written alternatively as

$$Y_i = \alpha + \sum_{j=1}^{p} \int X_{ij}(t)\beta_j(t)\, dt + \epsilon_i \tag{2}$$

where Y_i is the scalar response for observation i, $i = 1, \ldots, n$, and ϵ_i are independent random errors with mean 0 and variance σ^2, typically assumed to be normally distributed. For simplicity of exposition we present the main ideas for the continuous response case (2) first, and discuss the modifications required by a generalized response thereafter.

Current literature in generalized functional linear models is focused primarily on the estimation of the model components, the coefficient functions $\beta_j(\cdot)$. For example, Goldsmith et al. (2011a), James (2002), James et al. (2009), Marx & Eilers (1999), Müller & Stadtmüller (2005) and Tutz & Gertheiss (2010) discuss estimation and/or inference of the smooth effects functions $\beta_j(\cdot)$, in the case of one or multiple functional predictors. More recently, Kong et al. (2013) and Swihart et al. (2013), discuss hypothesis testing procedures when the number of functional predictors is assumed small. These methods perform well when the response variable is indeed related to most of the functional covariates. However, there is limited literature on the situation where the number of functional predictors available may be unnecessary large and many of them are unrelated to the response; see for example the data illustrations in Section 5. In these cases, typical association models that relate the response to all the measured predictors lead to unnecessary complex models that do not accurately describe the data generating process, and have low predictive capabilities. The focus of this paper is to develop procedures that perform variable selection when the predictors are curves, as well as estimation of the corresponding smooth effects of the selected functional variables.

Variable selection, especially in multivariate statistics, has attracted considerable attention over the recent years. In this context, penalized likelihood methods have emerged as highly successful selection techniques in presence of high dimensional vector-valued predictors; see for example LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001), the adaptive LASSO (Zou, 2006) and OSCAR (Bondell & Reich, 2008). The general idea of the penalized approaches is to impose various constraints on the coefficients, such as using L^1 norm or pairwise L^∞ norm etc., which result in sets of coefficients becoming identically zero. However, extension of these ideas to the setting of functional predictors is not straightforward. Usually the complexity of variable selection problems in functional regression is much greater than the usual multivariate variable selection problems. To fix the ideas, let us consider the diffusion tensor imaging tractography data that we use as an illustrative example in Section 5.1. In this data set, there are five tracts and six different measurements recorded for each tract, resulting in 30 functional predictors. The complexity arises because (i) each functional covariate is measured with error, (ii) some of the covariates are measured on irregularly spaced grid points, and (iii) not all the covariates have the same domain. Thus any variable selection procedure intended for this type of data set needs to account for all the above mentioned issues.

A direct generalization of multivariate variable selection ideas to setting (2) would require two steps. First, represent the integrals $\int X_{ij}(t)\beta_j(t)\, dt \approx \sum_l X_{ij}(t_l)\beta_j(t_l)$ using Riemann integration techniques and reformulate the problem in the

typical linear form. The second step is to apply classical variable selection procedures with high dimensional "artificial" predictors, $\{X_{j,l} = X_j(t_l) : j, l\}$. The main difficulty is the expected high correlation among predictors obtained from evaluating a single functional predictor, say X_j at different time points t_l, say $\{X_{j,1}, X_{j,2}, \ldots\}$. The challenge can be bypassed in turn by using methods developed for group variable selection, including group LARS and group LASSO (see, for example, Yuan & Lin, 2006), which target highly-correlated scalar predictors. These penalties, however, do not impose smoothness of the coefficients $\beta_{j,l} = \beta_j(t_l)$, viewed as functions of t. Without the smoothing property, interpretation of the functional predictors' influence on the response is meaningless.

In this article, we consider a penalized likelihood approach that combines selection of the functional predictors and estimation of the smooth effects for the chosen subset of predictors. Specifically, for a functional regression model (2), the estimates of the coefficient functions $\beta_j(\cdot)$ are the minimizers of $\sum_{i=1}^{n}\{Y_i - \alpha - \sum_{j=1}^{p}\int X_{ij}(t)\beta_j(t)\,dt\}^2 + \sum_{j=1}^{p} P_{\lambda,\varphi}(\beta_j)$, where $P_{\lambda,\varphi}$ is a penalty on the coefficient function β_j and λ and φ are global non-negative penalization parameters that control both the sparseness and the smoothness of the solution, respectively. More generally, when the response is generalized, the penalized criterion is similar to the above, with the difference that the first term is replaced by minus twice the corresponding (log-)likelihood function. We propose a penalty function that is inspired from the penalization proposed in high-dimensional additive models (Meier et al., 2009). Thus the proposed approach combines functional data analysis and variable selection techniques in high-dimensional additive models. Recently Fan & James (2012) (unpublished manuscript) investigated variable selection in linear and non-linear functional regression for continuous scalar response. They use a 'basis approach' modeling of the coefficient functions $\beta_j(\cdot)$, where the number of basis functions gives the smoothness of $\beta_j(\cdot)$, and is a tuning parameter, and a penalty term constructed on the norm of the entire effect, $\int X_j(t)\beta_j(t)\,dt$. In contrast, in this article, we propose a 'penalization approach' by modeling the coefficient functions using a rich function basis, and use a penalty term that explicitly controls the smoothness of $\beta_j(t)$ and sparseness of the model. Plus, our procedure is also developed for generalized functional linear models.

The remainder of the paper is organized as follows. Section 2 presents the methodology for simultaneous variable selection and estimation of smooth effects, when the predictors are functions observed without noise at a dense grid of points. Section 3 presents the extension to functional predictors corrupted by noise, measured at dense or sparse grid of points. The proposed methods are illustrated numerically in simulations experiments in Section 4 and on two real data examples in Section 5: the sugar spectra and tractography data sets. Section 6 concludes with a brief discussion.

2. Methodology

2.1. Parameters Modeling

Our approach to estimating the coefficient functions β_j is to use a pre-set basis functions expansion, such as a B-spline basis or a truncated polynomial basis. Specifically, let $\mathbf{b}_j(t) = \{b_{j1}(t), \ldots, b_{jq}(t)\}$ be such a finite basis, and consider the approximation as $\beta_j(t) = \sum_{r=1}^{q}\gamma_{jr}b_{jr}(t)$, where γ_{jr} are the corresponding basis coefficients. The choice of the basis functions is related to characteristics of the coefficient functions β_j, such as differentiability, while the number of basis functions is related to the coefficient's smoothness. In particular a small number of basis functions lead to a very smooth solutions, while a large number of basis functions results in very wiggly solutions. In this paper we allow the number of basis functions to be large enough to capture the complexity of the function, and we control the smoothness of the fit using an additional parameter that penalizes the roughness of the fit, and is chosen data-adaptively.

When the functional predictors, $X_{ij}(\cdot)$, are observed without measurement error and at an equally spaced dense grid of points, $\{t_{j1}, \ldots, t_{jN_j}\}$, the integral in (1) or (2) can be computed/approximated by the Riemann sum

$$\int X_{ij}(t)\beta_j(t)\,dt \quad \approx \quad \sum_r \left\{\Delta_j \sum_l X_{ij}(t_{jl})b_{jr}(t_{jl})\right\}\gamma_{jr} = Z_{ij}^{\top}\gamma_j,$$

where $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jq})^{\top}$, $Z_{ij} = (Z_{ij1}, \ldots, Z_{ijq})^{\top}$, $Z_{ijr} = \Delta_j \sum_l X_{ij}(t_{jl})b_{jr}(t_{jl})$, and $\Delta_j = t_{jl} - t_{j,l-1}$ denotes the

distance between two adjacent measurement points. Thus, model (2) can be approximated by a typical linear regression model $Y_i = \alpha + \sum_{j=1}^{p} Z_{ij}^{\top}\gamma_j + \epsilon_i$, where Z_{ij} are known quantities, and α and γ_j's are unknown regression coefficients. Nevertheless, using a classical penalty function on the model parameters γ_{jr}, in particular a sparseness inducing penalty, is not directly applicable to this setting because the parameters γ_{jr} have a preset grouping structure. Group variable selection, on the other hand, is an attractive method to impose sparseness. However, it implicitly assumes that the smoothness of the coefficient functions β_j is specifically controlled by the number of basis functions. For example, Matsui & Konishi (2011) and Lian (2013) discussed a groupwise SCAD penalty, by employing Gaussian basis functions, or the truncated eigenbasis of functional covariates, respectively, and controlling the smoothness of the corresponding functional coefficient by using only a small number of these basis functions; Zhu & Cox (2009) used a simple group lasso penalty and a very small number of orthonormal basis functions. When using only a small number of basis functions, however, the shape of the fitted functions is strongly influenced be the concrete number, the type and the placing of the basis functions. Our paper relaxes this limitation, and allows to approximate the coefficient functions using a large number of basis functions. This makes the approach more flexible, but does not fully describe the functions' inherent smoothness. Therefore a different type of penalty is required.

2.2. Penalized Estimation

We consider a so called *sparsity-smoothness penalty* technique, which results in simultaneous estimation of the parameter functions and sparseness of the solution, when the covariates are curves. The proposed penalty function was introduced by Meier et al. (2009) for variable selection in high-dimensional additive modeling. Specifically, let

$$P_{\lambda,\varphi}(\beta_j) = \lambda(\|\beta_j\|^2 + \varphi\|\beta_j''\|^2)^{1/2}, \tag{3}$$

where $\|\beta_j\|^2 = \int_{\mathcal{D}_j}\{\beta_j(t)\}^2 dt$ is the L^2 norm of β_j and $\beta_j''(t) = \partial^2\beta_j(t)/\partial t^2$. Penalties such as (3) ensure that estimates of the quantity appearing inside the square-root may be set to zero (see, e.g., Yuan & Lin (2006), Meier et al. (2008)). As $\varphi \geq 0$, this is equivalent to setting $\beta_j(t) = 0$ for all t, which implies that predictor X_j is excluded from model (1). If $\varphi > 0$, also the second order derivative of $\beta_j(t)$ is penalized, which ensures smoothing of non-zero coefficient functions. Although $\beta_j(t)$ is assumed as a smooth function, fitted non-zero curves may become rough if the penalty $\|\beta_j''\|^2$ on the second order derivative is not included in (3). As rough coefficient functions are hard to interpret, $\varphi > 0$ is strongly recommended. The actual extent of smoothing is controlled by the actual value of φ.

Using a rich pre-set basis function expansion for β_j, the corresponding penalty can be represented as

$$P_{\lambda,\varphi}(\beta_j) = \lambda(\gamma_j^{\top}(\Psi_j + \varphi\Omega_j)\gamma_j)^{1/2},$$

where Ψ_j is the $q \times q$ matrix with the (r,k) element equal to $(\Psi_j)_{rk} = \int b_{jr}(t)b_{jk}(t)\,dt$, $r,k = 1,\ldots,q$, and Ω_j is the $q \times q$ matrix with the (r,k) element equal to $(\Omega_j)_{rk} = \int b_{jr}''(t)b_{jk}''(t)\,dt$, $r,k = 1,\ldots,q$. Furthermore, the penalty can be re-written in a more convenient way as $P_{\lambda,\varphi}(\beta_j) = \lambda(\gamma_j^{\top}K_{\varphi,j}\gamma_j)^{1/2}$, which is a general group lasso-type penalty (Yuan & Lin, 2006) on vector γ_j, where $K_{\varphi,j} = \Psi_j + \varphi\Omega_j$ is a symmetric and positive definite $q \times q$ matrix.

Consequently, when the functional linear model (2) is assumed, the estimates of the intercept α and the coefficient functions β_j are obtained as $\widehat{\alpha}$ and $\widehat{\beta}_j(t) = \sum_{r=1}^{q} b_{jr}(t)\widehat{\gamma}_{jr}$, for $j = 1,\ldots,p$, where $\widehat{\alpha}$ and $\hat{\gamma}_j$'s are the minimizers of

$$\textstyle\sum_{i=1}^{p}(Y_i - \alpha - Z_{ij}^{\top}\gamma_j)^2 + \lambda\sum_{j=1}^{p}(\gamma_j^{\top}K_{\varphi,j}\gamma_j)^{1/2}, \tag{4}$$

where λ is a sparseness/tuning parameter and φ is a smoothing/tuning parameter. If $K_{\varphi,j} = I_q$ for all $j = 1,\ldots,p$, where I_q denotes the $q \times q$ identity matrix, then the estimation (4) resembles the one for the ordinary group lasso for which built in software already exist. More generally, by using appropriate re-parametrization one can still employ existing software, as we show next; the ideas were also pointed out by Meier et al. (2009) in a related context.

Let $K_{\varphi,j} = R_{\varphi,j}R_{\varphi,j}^\top$ be the Cholesky decomposition where $R_{\varphi,j}$ is non-singular lower triangular matrix and define $\tilde{\gamma}_j = R_{\varphi,j}^\top \gamma_j$ and $\widetilde{Z}_j = R_{\varphi,j}^{-1} Z_j$. Then, the penalized criterion (4) reduces to $\sum_{i=1}^p (Y_i - \alpha - \widetilde{Z}_{ij}^\top \tilde{\gamma}_j)^2 + \lambda \sum_{j=1}^p \|\tilde{\gamma}_j\|$, where $\|\tilde{\gamma}_j\|$ is the Euclidean norm in $\mathbb{R}^q$. Thus, for a given value of the smoothness parameter φ, the minimizers of (4) can be formulated as the parameter estimates in an appropriate linear model, via a penalized likelihood criterion using the ordinary group lasso penalty (Meier et al., 2008; Meier, 2009). As a result, for fixed φ and λ, the estimates $\hat{\tilde{\gamma}}$ can be computed using any existing software that can accommodate a group lasso penalty, for example, the R package `grplasso`. In practice, of course, the appropriate sparseness and smoothing parameters λ and φ are unknown and need to be estimated from the data; for example by cross-validation. This is discussed in detail in Section 2.4.

When the response is generalized and a generalized functional linear model as in (1) is assumed, the estimation procedure is similar to the one above with the exception that the first quadratic term of (4) is replaced by the (log-)likelihood function corresponding to Y_i as specified by (1).

2.3. Adaptive Penalized Estimation

Next we discuss an alternative penalty function based on adaptive weights for the penalized estimation criterion. We define the adaptive penalization scheme, similar to the adaptive lasso (Zou, 2006), by introducing weights w_j and v_j in the penalty function (3). Specifically, the adaptive penalization approach uses the penalty

$$P_{\lambda,\varphi}(\beta_j) = \lambda(w_j\|\beta_j\|^2 + \varphi v_j \|\beta_j''\|^2)^{1/2}, \tag{5}$$

where the weights w_j and v_j are chosen in a data-adaptive way (Meier et al., 2009). The choice of weights is meant to reflect some subjectivity about the true parameter functions and to allow for different shrinkage and smoothness for the different covariates. One possibility for choosing the weights is to use initial parameters estimates, based on smoothing solely, but without using sparseness-assumptions. Consider a generalized functional linear model with multiple functional covariates, and denote by $\breve{\beta}_j$'s, the estimated coefficient functions β_j's, for example, by using the approach described in Goldsmith et al. (2011a) and implemented in the R package `refund` (Crainiceanu et al., 2012). Then, the adaptive weights can be defined as $w_j = 1/\|\breve{\beta}_j\|$ and $v_j = 1/\|\breve{\beta}_j''\|$. Adaptive estimation has been shown to reduce the number of false positives considerably in penalty-based variable selection; see Meier et al. (2009) or Gertheiss & Tutz (2012) to name a few. As illustrated by the simulation studies in Section 4, adaptive estimation with the choice of weights above typically leads to improved selection performance in the functional model, too. Given the computation of the initial estimates $\breve{\beta}_j$ is not time-consuming, the computational burden for the adaptive penalty (5) does not change very much compared with the standard (non-adaptive) penalty (3).

2.4. Choosing the tuning parameters

So far, the sparseness parameter λ and the smoothness parameter φ involved in the penalties (3) and (5) were assumed known. In practice, however, these parameters need to be selected. We consider K-fold cross-validation to select λ and φ. Specifically, the original sample is (randomly) divided into K roughly equal-sized subsamples. For each $k = 1, \ldots, K$, the kth subsample is retained as a validation data for evaluating the prediction error of the model, and the remaining $K-1$ subsamples are used a training data. The K estimates of prediction error are averaged; the criterion selects the values of the tuning parameters that minimize the overall prediction error. Although the optimal value of K is still an open problem, typical values of K used in the literature are 5 and 10 (see also Hastie et al. (2009)).

As a measure of prediction accuracy, we use the predictive deviance $D = -2\phi \sum_{i=1}^n (l_i(\hat{\mu}_i) - l_i(Y_i))$, with $l_i(\hat{\mu}_i)$ denoting the individual log-likelihood (for observation i) at the fitted mean value $\hat{\mu}_i$ and $l_i(Y_i)$ being the corresponding log-likelihood where μ_i is replaced by the observed value Y_i (i.e., the maximum likelihood achievable). For the functional linear model (2) with normal ϵ_i, the predictive deviance simplifies to the the sum of squared errors $\sum_i (Y_i - \hat{\mu}_i)^2$.

3. Extension to Noisy and Sparse Functional Predictors

In this section we extend the variable selection methodology to more realistic functional data models, where the observed functional predictors are in fact proxy realizations of the underlying smooth trajectories X_{ij}. If the observed predictors are denoted by W_{ij}, then we write $W_{ij}(t) = X_{ij}(t) + e_{ij}(t)$, where e_{ij} is a white noise process. We assume that X_{ij} have mean zero and are squared integrable functions on $\mathcal{D}_j$. Our methodology is not directly applicable to the proxy covariates, and requires preliminary analysis to de-noise and re-construct the underlying curves. The approach varies according to whether the sampling design at which the functional covariates are observed, is dense or sparse.

First we focus on the setting when the functional predictors are observed on a dense grid of time points, $\{t_{ijl} : l\}$ in $\mathcal{D}_j$ for every i. When functional data are dense, then smoothing techniques can be applied to each curve in turn, to reconstruct the de-noised trajectories, $\widehat{X}_{ij}$. For example (global) spline smoothing is used in Ramsay & Silverman (2005), while local polynomial Kernel smoothing is used in Zhang & Chen (2007). In particular, for local polynomial smoothing Zhang & Chen (2007) show that, under mild conditions, the underlying smooth curves are reconstructed with asymptotically negligible error. The main challenge is the selection of the smoothing parameter, which could be either curve-specific, or constant across all the curves. Zhang & Chen (2007) suggest a generalized cross validation criteria (GCV) using a global smoothing parameter; another criterion is the restricted maximum likelihood (REML).

Consider next the setting when the functional predictors are observed on a sparse grid of time points, $\{t_{ijl} : l\}$, such that the set of all observation points $\{t_{ijl} : l, i\}$ is dense in $\mathcal{D}_j$. In this case, smoothing each curve is no longer an option as the number of observations per curve is typically very small. Instead, we use the functional principal component analysis (FPCA) technique, one of the main toolkit in functional data analysis. This technique is based on three steps, see e.g., Yao et al. (2005), and as we briefly review it here. First the data $\{W_{ij}(t_{ijl}) : l\}_i$ are pooled together, to obtain smooth estimates of the mean function $\mu_j(t) = E[X_{ij}(t)]$ and of the covariance function $K_j(t, t') = \text{cov}\{X_{ij}(t), X_{ij}(t')\}$; recall it is assumed that $\mu_j \equiv 0$. Second the spectral decomposition of the estimated covariance function yields a finite number of pairs $\{\widehat{\phi}_{j,\ell}(t), \widehat{\lambda}_{j,\ell}\}_\ell$ of estimated eigenfunctions and corresponding eigenvalues, with $\widehat{\lambda}_{j,1} > \widehat{\lambda}_{j,2} > \ldots > 0$. Third, the underlying curves $X_{ij}(t)$ are estimated using a finite eigenbasis function truncation, $\widehat{X}_{ij}(t) = \sum_{\ell=1}^{N_j} \widehat{\xi}_{j,1\ell}\widehat{\phi}_{j,\ell}(t)$, where $\widehat{\xi}_{j,1\ell}$ are obtained using conditional expectation; see Yao et al. (2005). There are multiple ways to select the finite truncation available in literature, such as AIC, BIC etc.; from our empirical experience the simple criterion based on percentage of explained variance (such as 90% or 95%) gives satisfactory results. For carrying out FPCA, we use `fpca.sc()` (Di et al., 2009) from the R package `refund` (Crainiceanu et al., 2012).

Once the underlying curves are estimated, and evaluated on a dense grid of points, then the variable selection methodology proceeds as detailed in Section 2, by pretending that the estimates are the true functional covariates.

4. Numerical Experiments

4.1. An Illustrative Toy Example

Consider an example in which two functional covariates are observed at a set of 300 equidistant points in $(0, 300)$ for each sampling unit. Define (similar to Tutz & Gertheiss (2010)) for $i = 1, \ldots, 300$, and $j = 1, 2$,

$$X_{ij}(t) = \{\sigma(t)\}^{-1}\sum_{r=1}^{5}(a_{ijr}\sin(\pi t(5 - a_{ijr})/150) - m_{ijr}), \tag{6}$$

where $t \in \mathcal{D} = (0, 300)$, $a_{ijr} \sim U(0, 5)$, $m_{ijr} \sim U(0, 2\pi)$, with $U(a, b)$ denoting the uniform distribution on interval $[a, b]$. Here $\sigma(t)$ is defined such that $var\{X_{ij}(t)\} = 0.01$ for all $t \in (0, 300)$. For the response, we assume a functional linear model $Y_i = \alpha + \int_0^{300}\beta_1(t)X_{i1}(t)dt + \epsilon_i$, where $\epsilon_i \sim N(0, 2^2)$, and $\beta_1(t)$ has a Gamma density-like shape (see the gray curve in Figure 1). Thus, the response depends only on the first functional covariate and not on the second.

Figure 1 depicts the fitted coefficient functions obtained by the proposed functional variable selection (3) for different values of the sparseness parameter λ and the smoothing parameter φ. Following our intuition, as the sparseness parameter λ increases, the estimated coefficient functions are shrunk and at some value, set to zero (see in particular the solid curves in Figure 1, which refer to very large $\lambda = 10^{10}$). The exact λ-value making the estimated coefficient function vanish, depends also on the chosen φ. More importantly, φ influences the smoothness of the fitted functions. As the smoothing parameter φ increases, the departure from linearity is penalized stronger and thus the estimated curves become closer to a linear function. Smaller values for φ, are related to low penalization of the departures from linearity, and thus result in very wiggly and difficult to interpret estimated coefficient functions. For optimal estimates (in terms of accuracy and interpretability), an adequate (λ, φ) combination has to be chosen.

This provides evidence that when the covariate are functions, smoothing is absolutely necessary. A simple group lasso-type penalty, or other similar group selection penalties placed on sets of basis coefficients, without controlling for smoothness are not viable solutions for functional variable selection.

4.2. Simulation Study

We conducted simulations in a variety of settings to illustrate the performance of the proposed method in terms of sparseness accuracy and prediction performance. In this section, we summarize the results based on data sets of the form $[\{X_{i1}(t) : t \in \mathcal{T}_1\}, \ldots, \{X_{i10}(t) : t \in \mathcal{T}_{10}\}, Y_i]_i$, where $\mathcal{T}_j$ is the set of points at which the covariate X_{ij} is observed and $i = 1, \ldots, 300$. The functional predictors are defined as in (6). A variety of settings are obtained by combining:

Scenario A Two sampling designs are considered for the functional covariates:

(i) Dense sampling design, where $\mathcal{T}_j$ is taken to be the set of 300 equidistant points in $(0, 300)$.
(ii) Moderately sparse and irregularly sampled design points, where the size $\mathcal{T}_j$ is uniformly generated between 20 and 30, and the set of time points is uniformly sampled from $\mathcal{T}$ – the set of 300 equidistant points in $(0, 300)$.

Scenario B Two generating distributions are considered for the response:

(i) Functional Linear Model, and in particular $Y_i = \alpha + \sum_{j=1}^{5} \int_0^{300} \beta_j(t) X_{ij}(t) dt + \epsilon_i$, where $\epsilon_i \sim N(0, 10)$, and $\beta_1(t), \beta_2(t), \beta_3(t)$ have Gamma-density like shape with effect sizes decreasing with increasing j, and $\beta_4(t)$ and $\beta_5(t)$ have exponential like shape with $\beta_5(t)$ being more linear (see Figure 2, top row).
(ii) Generalized Functional Linear Model, and in particular $Y_i \sim \text{Bernoulli}[\exp(\eta_i)/\{1 + \exp(\eta_i)\}]$ and $\eta_i = \alpha + \sum_{j=1}^{5} \int_0^{300} \beta_j(t) X_{ij}(t) dt$. Regression coefficients from (i) are scaled such that true success probabilities are rather uniformly distributed on $[0, 1]$, with maximum probability mass around 0.1 and 0.9.

Note, only signals $j = 1, \ldots, 5$ are assumed to be relevant (with true coefficient functions as described in B(i)). The error variance $\sigma^2 = 10$ for the linear model B(i) is chosen such that the signal to noise ratio is comparable to the one found in the real data application in Section 5.2.

To assess the performance of the proposed method two alternative approaches are considered. The first takes the 'basis approach' modeling for smooth effects (see also Fan & James (2012)) using 30 basis functions and then employs our method with penalty (3) for $\varphi = 0$, and uses the number of basis functions to implicitly control the smoothness of the effects (Simple). The second is the penalized functional regression (PFR; Goldsmith et al., 2011a), which imposes smoothing and selects the smoothing parameters by REML, but does not do any selection of the variables; this method is implemented in the `pfr()` function of the R package `refund` (Crainiceanu et al., 2012). Three versions of the proposed methodology are examined: the standard functional group lasso (Standard), the adaptive version (5), with both w_j and v_j chosen adaptively (Adapt1), and the semi-adaptive version with adaptively chosen w_j and fixed $v_j = 1$ (Adapt2). The initial estimates for the adaptive methods (Adapt1 and Adapt2) are obtained using PFR.

Figure 2 presents boxplots of the squared errors (SE) observed in 50 independent simulation runs, for the different methods, where $SE = \int(\hat{\beta}_j(t) - \beta_j(t))^2 dt$, $\beta_j(t)$ and $\hat{\beta}_j(t)$ are the true and estimated coefficient functions, respectively. The results show that, for both functional linear and generalized functional linear model, and irrespective of the sampling design, the proposed methods Standard and Adapt2 give the highest accuracy (depicted in Figure 2 by red/blue colors). Adapt1 (green color) performs slightly worse, and this is due to the fact that the initial estimates with PFR are over-smooth. The intuition is that in the case of over-smoothness, the adaptive weights v_j, which are defined as the reciprocal of the smoothness measure, are very large, causing instability in the selection of the tuning parameters. When $\varphi = 0$ (Simple) or $v_j = 1$ (Adapt1, Standard) the criterion yields more reliable estimates for the tuning parameter(s), which explains why these methods perform well. However, the Simple method (turquoise color) yields under-smooth effects, hardly interpretable. As expected, the PFR performs worst for all the models and designs.

To investigate the predictive capabilities of the proposed methods, we generated a test data set with $m = 5000$ observations. Figure 3 (left), shows the resulting mean squared errors of prediction $m^{-1}\sum_{i=1}^{m}(Y_i - \hat{Y}_i)^2$ for the five methods. Overall, Adapt2 (blue color) has the best performance, and moreover all the variable selection methods have a superior performance to the PFR. For the logistic model the prediction performance is defined using the predictive log score $m^{-1}\sum_{i=1}^{m}\{Y_i \log(1/\hat{\pi}_i) + (1 - Y_i)\log(1/(1 - \hat{\pi}_i))\}$, since the log score is proportional to our deviance criterion; here $\hat{\pi}_i$ denotes the estimated probability of $Y_i = 1$. The results are consistent across the different sampling designs (dense/sparse) reinforcing that reconstructing the functional covariates by FPCA performs very well.

We further investigate the methods in terms of how frequent they select the true model; only the four variable selection methods are compared. Table 1 gives the proportion of simulation runs (linear model, dense design) for which each functional predictors is selected, corresponding to every method. It shows that the percentage of false positives (predictors 6–10) is considerably reduced for the adaptive versions of our method. While estimation accuracy was a little lower for the fully adaptive method (Adapt1), selection performance in terms of false positive rates is slightly better than with adaptive w_j only (Adapt2). The relatively high false positive rates (around 40%), compared to very low false negatives, could be due to the tendency of our criterion, inherited from cross-validation, to select accurate estimates but a somewhat larger model.

5. Real Data Applications

5.1. Tractography Data

Our motivating application is a neurological study of the white matter tracts in the brains of multiple sclerosis (MS) patients and healthy subjects using magnetic resonance imaging (MRI) techniques; the study has been previously described in Greven et al. (2010), Staicu et al. (2012), Goldsmith et al. (2012), and Gertheiss et al. (2013). MS is a neurological disease that affects the central nervous system and in particular damages white matter tracts in the brain through lesions, myelin loss and axonal damage. Modern MRI allows the extraction of information on individual tracts and thus allows a better understanding of damages in neuronal tracts. In particular diffusion tensor imaging (DTI) has been successfully used for examining white matter tracts through modalities that measure water diffusivity along the tracts; see, for example, Basser et al. (1994), Basser et al. (2000). Six MRI indices – T2 relaxation time (T2), magnetization transfer ratio (MTR), mean diffusivity (MD), fraction anisotropy (FA), parallel diffusivity (L0) and radial/perpendicular diffusivity (LT) – are obtained at many locations along the tract, for five well identified white matter tracts: right/left corticospinal tract (CST), corpus callosum (CC) and right/left optic radiations tract (OR). Due to the complexity of MS, it is not known which MRI indices are related to the disease.

Previous literature focused mainly on using one or two MRI indices along a specified white matter tract to predict the disease status. For example Goldsmith et al. (2011a) used L0 along the CC tract, McLean et al. (2013) used either

FA or L0 along the CC, and Goldsmith et al. (2012) uses MD along the CC tract and LT along the right CST in a longitudinal analysis. However, as Reich et al. (2007) pointed out, 'any single MRI index has a unique pathological correlate' and thus 'the ability to examine multiple MRI indices simultaneously is an important step in acquiring a more complete description of the damage done to brain tissue by neurological disease'. Goldsmith et al. (2011b) attempted to use all the MRI indices along CC, CST and OR tracts for predicting the disease status; however, to bypass the very large dimensionality of the predictors and the limited sample size, the authors have combined, for each of CST and OR, the corresponding MRI indices along the left and right tracts. However, the MRI indices along the same tract are correlated, and thus a full analysis, may imply an unnecessarily complex model. In this paper, we consider all the MRI indices along the five different tracts, and study simultaneously (1) which of the profiles are important in predicting the disease status, and (2) what is their corresponding effect on the risk of developing the disease.

The data contain tract profiles for 168 MS subjects and 43 healthy volunteers observed at their baseline visit. The profiles of the various MRI indices have different 'domains' according to the tract along which they are measured. Many of the profiles include (varying degrees of) missingness, and all profiles are measured with error. Thus at a preliminary step we use the FPCA methodology as described in Section 3 to reconstruct the smooth profiles. The tract profiles are then de-meaned and scaled appropriately to have zero mean and unit variance at each location along the tract. We use a Bernoulli model with logistic link to relate the binary disease outcome to the profile covariates. Then, we apply the proposed functional variable selection technique, where the penalty parameters are chosen via cross-validation.

Figure 5 shows the results: MRI indices profiles that are selected to be important for disease prediction correspond to estimated smooth effects that are non-zero, while profiles deemed to be not important correspond to estimated effects that are zero. A number of nine tract profiles are chosen to be predictive of the disease status and their estimated effect is illustrated in Figure 5. For example, L0 above population average along the left OR tract indicates controls, as the corresponding coefficient function has negative values. High values of LT along the CC, by contrast, yields higher probability for being an MS patient, as this coefficient function has positive values. These findings make good sense, as L0 measures diffusivity along the main axis of the tract and LT measures diffusivity in perpendicular direction, with the first indicating healthy tracts and the latter indicating damage to the tract. Furthermore, we see that MD is completely excluded from the model, as all corresponding coefficient functions (3rd column in Figure 5) are zero. One possible explanation is due to the fact that MD is proportional to overall diffusivity (i.e., diffusivity in any direction) and hence less useful for indicating damage to the tract of interest. As different tracts/measures are of course dependent and lasso-type approaches typically select just one predictor from a group of highly correlated covariates, there may be alternative sets of functional predictors that are useful for classification of disease status. The proposed approach, however, provided interesting tract/measure combinations, which may be further investigated in subsequent analyses.

5.2. Sugar Spectra

Our second application comes from chemometrics, which is becoming a typical field for functional data analysis. In chemometrics there are often function-like absorbance or emission spectra given – in particular for food samples – that are used to determine the content of certain ingredients. Using the spectra is typically much cheaper than alternative chemical analysis. Traditionally, techniques like principal components regression (PCR) and partial least squares (PLS) are used in chemometrics, but recently functional data analysis tools have gained more and more attention.

We consider a data set described by Bro et al. (1998) and Bro (1999): 268 samples of sugar were dissolved and the solution was measured spectrofluorometrically. For every sample the emission spectra from 275–560 nm were measured in 0.5 nm intervals (i.e., at 571 wavelengths) at seven excitation wavelengths: 230, 240, 255, 290, 305, 325, and 340 nm. In addition, there are laboratory determinations of the quality of the sugar given, such as ash content (in percentage). Ash content measures the amount of inorganic impurities in the refined sugar, cf. Bro (1999). The aim of the analysis is to study the association between the ash content and the fluorescence spectra. As already pointed

out above, using the emission spectra is much easier and cheaper then chemical analysis in the lab, and the analysis would become even easier, if not all seven excitation wavelength had to be used.

In contrast to the tractography data, the response (ash content) is continuous. We use the functional linear model (2) with our approach to determine the most useful excitation wavelength. As variability of the spectra is very different for different excitation wavelengths, curves are standardized before applying our method. Figure 4 shows the estimated coefficient functions when using the standard penalty (3) (dashed red) or the adaptive version (5) with adaptive w_j (solid black). Tuning parameters were chosen via five-fold cross-validation. We see that the results are very similar for both methods. The major difference is that using the adaptive penalty not only excitation wavelength 230 nm and 255 nm are excluded but also 305 nm. Predictions of ash content are very similar (and very good) for the two models. The ratio of prediction error $\sum_i (Y_i - \hat{\mu}_i)^2$ and overall variation $\sum_i (Y_i - \bar{Y})^2$ is 15.2% for the non-adaptive penalty and 14.9% for the adaptive one. As the model obtained with the adaptive penalty is sparser while producing slightly better predictions, we prefer the solid black coefficient curves in Figure 4 produced by the adaptive version with adaptive w_j's.

6. Summary and Discussion

We proposed a variable selection procedure for generalized functional linear regression models where multiple functional predictors are present but only a few of these predictors are actually useful in predicting the response. Typical estimation procedures for such models do not consider the issue of selecting the useful predictors, and thus may suffer from overly large models and reduced predictive capability. Our procedure simultaneously selects the important functional variables, and estimates the corresponding effects. The smoothness of the coefficient functions and sparseness of the model are simultaneously controlled using a sparsity-smoothness penalty technique that controls roughness of the original coefficient functions as well as their second derivatives. R-code implementing the procedure is available as supplementary material.

We also investigated two adaptive versions of this approach where the penalty term is weighted using adaptively chosen weights. We found that our proposed methods perform well in terms of prediction error as well as mean squared errors for the estimated coefficient functions compared to fitting a model without any selection. In terms of selecting the correct predictors and number of false positives, adaptive methods perform better than the non-adaptive one while retaining reasonable estimation accuracy. Also, our method can be applied even when the functional predictors are observed with measurement error and on a sparse set of points.

As in any variable selection procedure, our method also requires estimation of the tuning parameters. Specifically, we require to estimate two tuning parameters: one overall tuning parameter and one smoothness parameter. This is done by K-fold cross-validation in this paper. We recognize that there are other approaches to choose such parameters and cross-validation may not be theoretically the best procedure. However, to the best of our knowledge, this issue is an open problem, especially in the framework of functional regression. While cross-validation performs well in our numerical studies, more research is needed to investigate the optimality of such estimates of the tuning parameters.

Acknowledgement

A. Maity's work was partially NIH grant R00ES017744 and an award from NCSU Faculty Research & Professional Development Fund (2012-2704). A.-M. Staicu's research was partly supported by NSF grant number DMS 1007466 and an award from NCSU Faculty Research & Professional Development Fund (2012-2673). The Content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Health. We particularly thank Ciprian Crainiceanu, Daniel Reich, the National Multiple Sclerosis Society, and Peter Calabresi for providing the tractography data and related information.

References

Basser, PJ, Mattiello, J & Le Bihan, D (1994), 'MR diffusion tensor spectroscopy and imaging,' *Biophysical Journal*, **66**, pp. 259–267.

Basser, PJ, Pajevic, S, Pierpaoli, C, Duda, J & Aldroubi, A (2000), 'In vivo fiber tractography using DT-MRI data,' *Magnetic Resonance in Medicine*, **44**, pp. 625–632.

Bondell, HD & Reich, BJ (2008), 'Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR,' *Biometrics*, **64**, pp. 115–123.

Bro, L, Nørgaard, L, Engelsen, SB & an C. A. Andersson, RB (1998), 'Chemometrics in food science a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance,' *Chemometrics and Intelligent Laboratory Systems*, **44**, pp. 31–60.

Bro, R (1999), 'Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis,' *Chemometrics and Intelligent Laboratory Systems*, **46**, pp. 133–147.

Crainiceanu, CM, Reiss, P, Goldsmith, J, Huang, L, Huo, L, Scheipl, F, Greven, S, Harezlak, J, Kundu, MG & Zhao, Y (2012), *refund: Regression with Functional Data*, R package version 0.1-6.

Di, C, Crainiceanu, C, Caffo, B & Punjabi, N (2009), 'Multilevel functional principal component analysis,' *Annals of Applied Statistics*, **3**, pp. 458–488.

Fan, J & Li, R (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties,' *Journal of the American Statistical Association*, **96**, pp. 1348–1360.

Fan, Y & James, GM (2012), 'Functional additive regression,' *Preprint*.

Gertheiss, J, Goldsmith, J, Crainiceanu, C & Greven, S (2013), 'Longitudinal scalar-on-functions regression with application to tractography data,' *Biostatistics*, to appear, doi: 10.1093/biostatistics/kxs051.

Gertheiss, J & Tutz, G (2012), 'Regularization and model selection with categorial effect modifiers,' *Statistica Sinica*, **22**, pp. 957–982.

Goldsmith, J, Bobb, J, Crainiceanu, C, Caffo, B & Reich, D (2011a), 'Penalized functional regression,' *Journal of Computational and Graphical Statistics*, **20**, pp. 830–851.

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

Goldsmith, J, Crainiceanu, C, Caffo, B & Reich, D (2012), 'Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements,' *Journal of the Royal Statistical Society: Series C*, **61**, pp. 453–469.

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

Hastie, T, Tibshirani, R & Friedman, JH (2009), *The Elements of Statistical Learning*, Springer, New York, 2nd edn.

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

James, GM, Wang, J & Zhu, J (2009), 'Functional linear regression that's interpretable,' *The Annals of Statistics*, **37**, pp. 2083–2108.

Kong, D, Staicu, A & A.Maity (2013), 'Classical testing in functional linear models,' *Preprint*, submitted.

Lian, H (2013), 'Shrinkage estimation and selection for multiple functional regression,' *Statistica Sinica*, **23**, pp. 51–74.

Marx, BD & Eilers, PHC (1999), 'Genaralized linear regression on sampled signals and curves: A p-spline approach,' *Technometrics*, **41**, pp. 1–13.

Matsui, H & Konishi, S (2011), 'Variable selection for functional regression models via the L1 regularization,' *Computational Statistics and Data Analysis*, **55**, pp. 3304–3310.

McLean, MW, Hooker, G, Staicu, AM, Scheipl, F & Ruppert, D (2013), 'Functional generalized additive models,' *Journal of Computational and Graphical Statistics*, to appear.

Meier, L (2009), *grplasso: Fitting user specified models with Group Lasso penalty*, R package version 0.4-2.

Meier, L, Van de Geer, S & Bühlmann, P (2008), 'The group lasso for logistic regression,' *Journal of the Royal Statistical Society B*, **70**, pp. 53–71.

Meier, L, Van de Geer, S & Bühlmann, P (2009), 'High-dimensional additive modeling,' *The Annals of Statistics*, **37**, pp. 3779–3821.

Müller, HG & Stadtmüller, U (2005), 'Generalized functional linear models,' *The Annals of Statistics*, **33**, pp. 774–805.

Ramsay, JO & Dalzell, CJ (1991), 'Some tools for functional data analysis,' *J. Roy. Statist. Soc. Ser. B*, **53**(3), pp. 539–572, with discussion and a reply by the authors.

Ramsay, JO & Silverman, BW (2005), *Functional Data Analysis*, Springer, New York, 2nd edn.

Reich, D, Smith, S, Zackowski, K, Gordon-Lipkin, E, Jones, C, Farrell, J, Mori, S, van Zijl, P & Calabresi, P (2007), 'Multiparametric magnetic resonance imaging analysis of the corticospinal tract in multiple sclerosis,' *NeuroImage*, **38**, pp. 271–279.

Staicu, AM, Crainiceanu, C, Reich, D & Ruppert, D (2012), 'Modeling functional data with spatially heterogeneous shape characteristics,' *Biometrics*, **68**, pp. 331–343.

Swihart, B, Goldsmith, J & Crainiceanu, CM (2013), 'Testing for functional effects,' *Preprint*, submitted.

Tibshirani, R (1996), 'Regression shrinkage and selection via the lasso,' *Journal of the Royal Statistical Society B*, **58**, pp. 267–288.

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

Yao, F, Müller, HG & Wang, JL (2005), 'Functional data analysis for sparse longitudinal data,' *Journal of the American Statistical Association*, **100**, pp. 577–590.

Yuan, M & Lin, Y (2006), 'Model selection and estimation in regression with grouped variables,' *Journal of the Royal Statistical Society B*, **68**, pp. 49–67.

Zhang, JT & Chen, J (2007), 'Statistical inferences for functional data,' *The Annals of Statistics*, **35**, pp. 1051–1079.

Zhu, H & Cox, DD (2009), 'A functional generalized linear model with curve selection in cervical pre-cancer diagnosis using fluorescence spectroscopy,' *IMS Lecture Notes, Monograph Series – Optimality: The Third Erich L. Lehmann Symposium*, **57**, pp. 173–189.

Zou, H (2006), 'The adaptive lasso and its oracle properties,' *Journal of the American Statistical Association*, **101**, pp. 1418–1429.

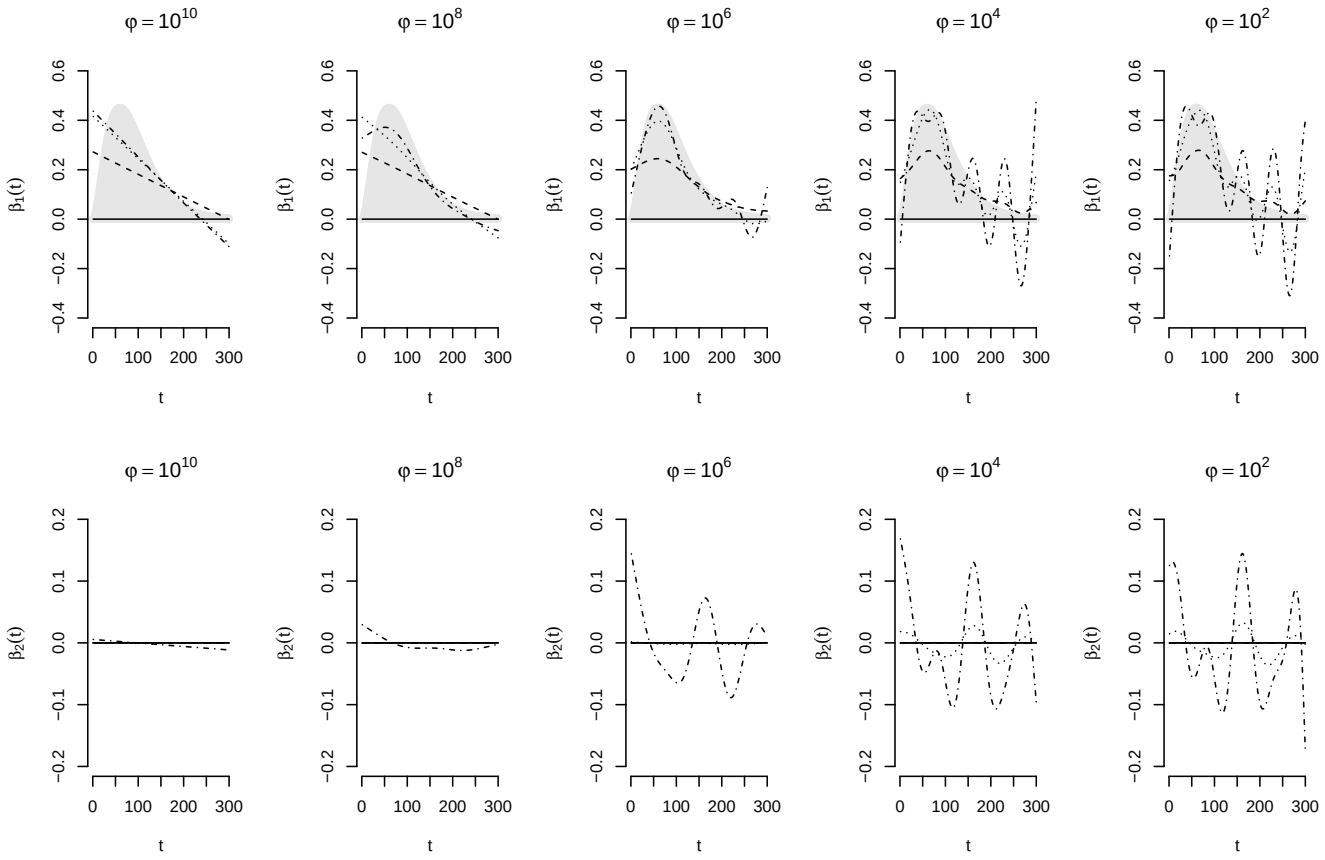

Figure 1. Fitting results for the functional group lasso for $\lambda \in \{10^3, 10^2, 10^1, 10^0\}$ (solid, dashed, dotted, dashed/dotted, respectively) and $\varphi \in \{10^{10}, 10^8, 10^6, 10^4, 10^2\}$ on a simulated dataset with two functional covariates but only the first one being relevant; the first row corresponds to the first covariate (with the true coefficient function shaded in grey), the second row to the actually irrelevant signals.

Table 1. Proportions of simulation runs with the respective functional predictor $j = 1, \ldots, 10$ being selected and average model size (dense design, normal response). Considered are the standard functional group lasso (Standard), the adaptive version with both w_j and v_j chosen adaptively (Adapt1), or adaptive w_j only (Adapt2), and the simple version without smoothing (Simple). Predictors 1–5 are truly relevant, covariates 6–10 are actually irrelevant.

	1	2	3	4	5	6	7	8	9	10	av. model size
Standard	1.00	1.00	0.98	1.00	1.00	0.72	0.74	0.72	0.78	0.68	8.62
Adapt1	1.00	1.00	0.90	1.00	1.00	0.44	0.42	0.38	0.40	0.42	6.96
Adapt2	1.00	1.00	0.90	1.00	1.00	0.48	0.44	0.46	0.46	0.38	7.12
Simple	1.00	1.00	0.98	1.00	1.00	0.72	0.78	0.78	0.80	0.66	8.72

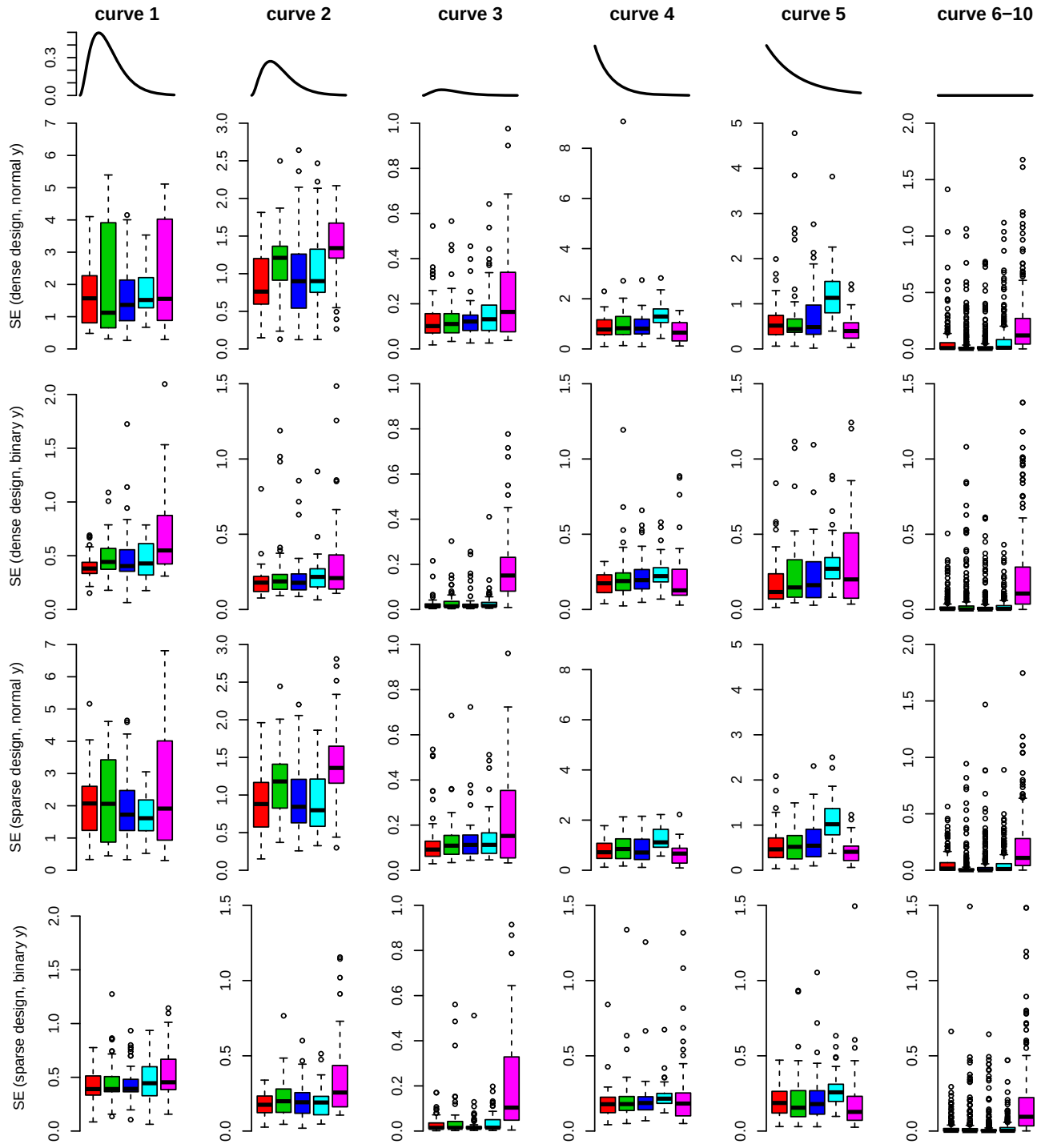

Figure 2. True coefficient functions (top row) and corresponding squared errors (SE) for the standard functional group lasso (red), the adaptive version with both w_j and v_j chosen adaptively (green), adaptive w_j only (blue), the simple version without smoothing (turquoise), and the pfr method without variable selection (purple); dense design/normal response (2nd row), dense design/binary response (3nd row), sparse design/normal response (4th row), sparse design/binary response (bottom row); for binary response a few outliers are not shown.

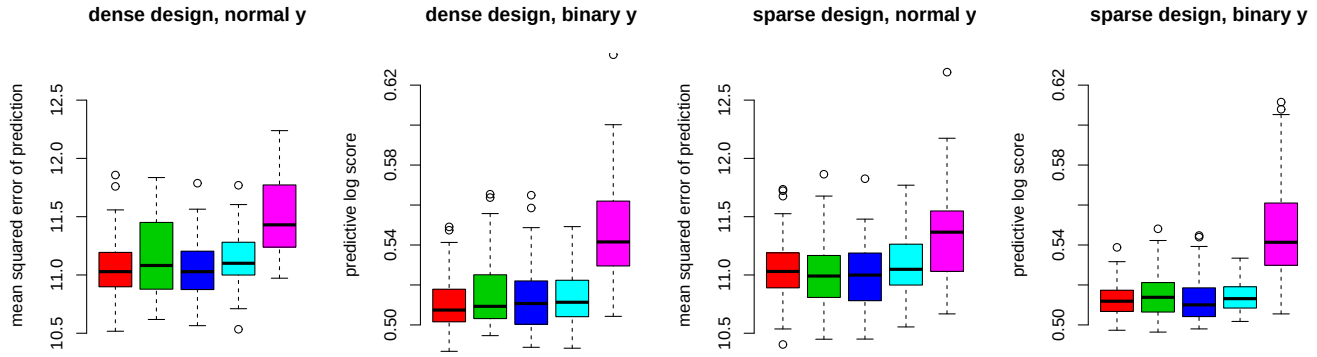

Figure 3. Mean squared errors of prediction/log scores on the test set for the standard functional group lasso (red), the adaptive version with both w_j and v_j chosen adaptively (green), adaptive w_j only (blue), the simple version without smoothing (turquoise), and the pfr method without variable selection (purple).

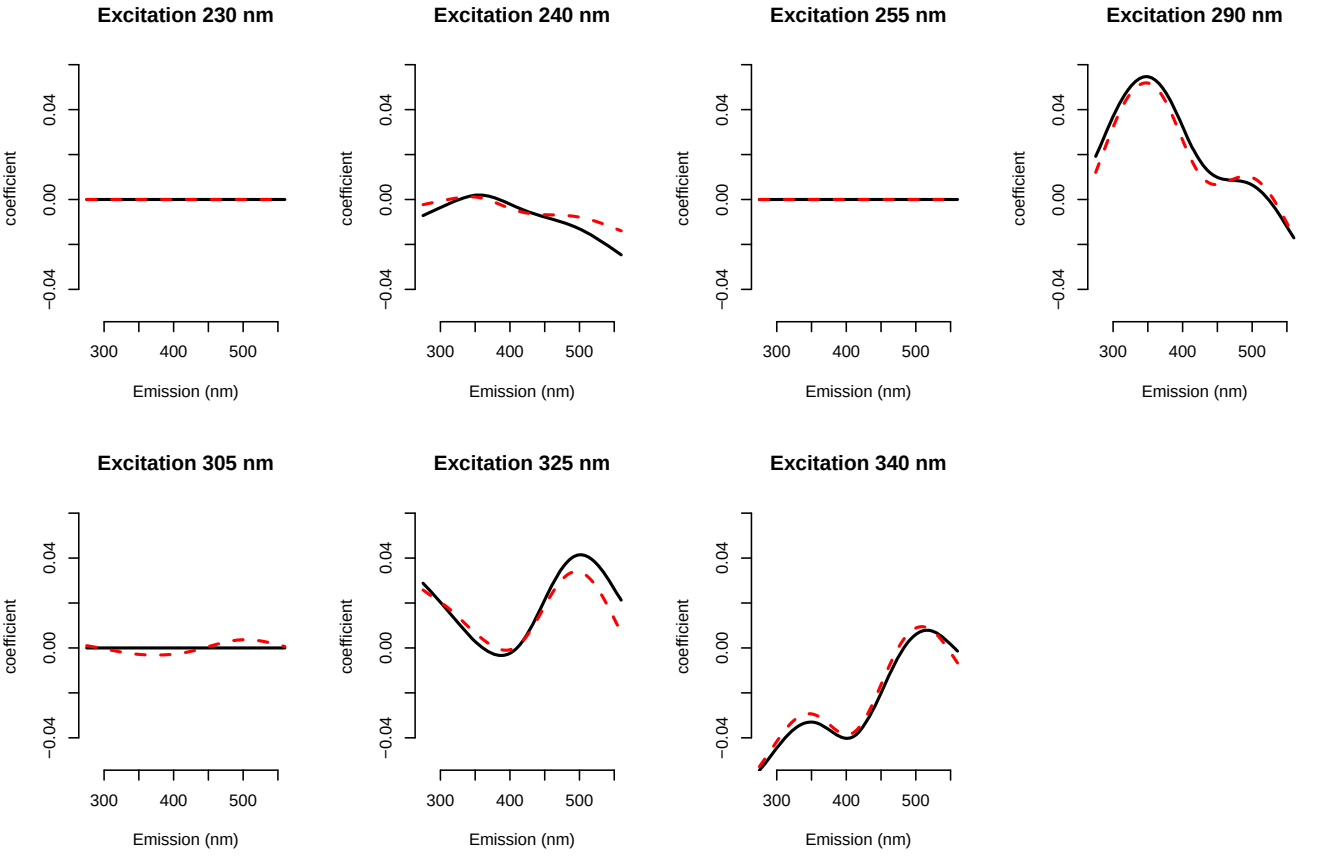

Figure 4. Estimated coefficient functions for the sugar data when using the standard functional group lasso with penalty (3) (dashed red) or the adaptive version (5) with adaptive w_j (solid black).

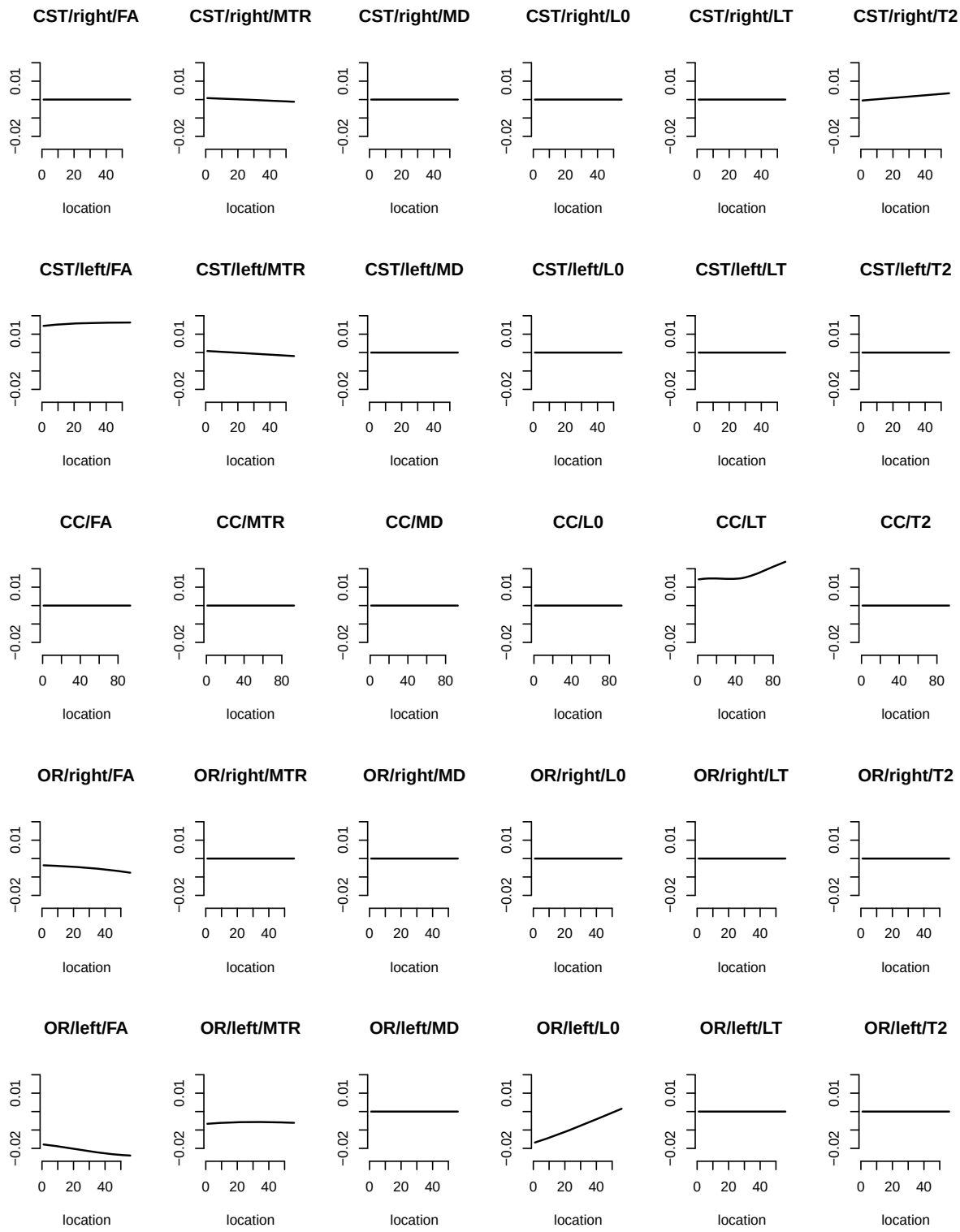

Figure 5. Estimated coefficient functions for the tractography data when using the standard functional group lasso with penalty (3).