Stat

The ISI's Journal for the Rapid Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.106

Variable selection in function-on-scalar regression

Yakuan Chen*, Jeff Goldsmith and R. Todd Ogden

Received 10 January 2016; Accepted 5 February 2016

For regression models with functional responses and scalar predictors, it is common for the number of predictors to be large. Despite this, few methods for variable selection exist for function-on-scalar models, and none account for the inherent correlation of residual curves in such models. By expanding the coefficient functions using a *B*-spline basis, we pose the function-on-scalar model as a multivariate regression problem. Spline coefficients are grouped within coefficient function, and group-minimax concave penalty is used for variable selection. We adapt techniques from generalized least squares to account for residual covariance by "pre-whitening" using an estimate of the covariance matrix and establish theoretical properties for the resulting estimator. We further develop an iterative algorithm that alternately updates the spline coefficients and covariance; simulation results indicate that this iterative algorithm often performs as well as pre-whitening using the true covariance and substantially outperforms methods that neglect the covariance structure. We apply our method to two-dimensional planar reaching motions in a study of the effects of stroke severity on motor control and find that our method provides lower prediction errors than competing methods. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: group MCP; kinematic data; pre-whitening; splines

1 Introduction

Regression models with functional responses and scalar predictors are routinely encountered in practice. These models face a challenge that also arises for traditional models: how to identify the important predictors among a potentially large collection. Functional response models face the additional challenges of high dimensionality and residual correlation. There are few methods for variable selection in this class of models, and none of them properly account for the correlation structure in the residuals. The purpose of this article is to address the current lack of methods for variable selection that accounts for residual correlation in function-on-scalar regression problems.

Our work is motivated by two-dimensional planar reaching data. As an assessment of upper extremity motor control, stroke patients and healthy controls made repeated reaching movements from a central point to eight targets arranged on a circle. The dataset consists of 57 subjects, including 33 patients suffering a unilateral stroke (meaning only one arm is affected) and 24 healthy controls, and contains motions made with both the dominant and non-dominant hands to each of the eight targets. Our analytic goal is to explore the effects of the potential predictors of motor control on these motions and to identify the most essential ones using variable selection. Among the potential predictors, the Fugl-Meyer score is a quantity that measures the severity of arm motor impairment (Fugl-Meyer et al., 1975). It ranges from 0 to 66 with smaller values indicating more severe impairment and 66 indicating healthy function.

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10027, USA
***Email: yc2641@cumc.columbia.edu**

(wileyonlinelibrary.com) DOI: 10.1002/sta4.106

Other potentially important predictors include target direction, whether the hand used was the dominant or non-dominant, and whether the hand used was contralesional (directly affected by the stroke) or ipsilesional (indirectly affected or unaffected).

Figure 1 shows the observed reaching motions for three subjects: a stroke patient with contralesional dominant hand in the left column; a stroke patient with contralesional non-dominant hand in the centre column and a heathy control in the right column. Reaching motions made by contralesional hand display deviation from straight paths from the starting point to each target; these deviations may be consistent for contralesional dominant or non-dominant hands. While deviation from straightness is not obvious in the ipsilesional arm, other effects, like over-reach, are observed. The potential for differential effects of stroke severity on reaching motions indicates the importance of allowing interactions between predictors of interest.

The observed data are horizontal and vertical coordinates of the hand position for each reaching motion as functions of time. We construct function-on-scalar regression models for the two outcome functions separately. Given scalar

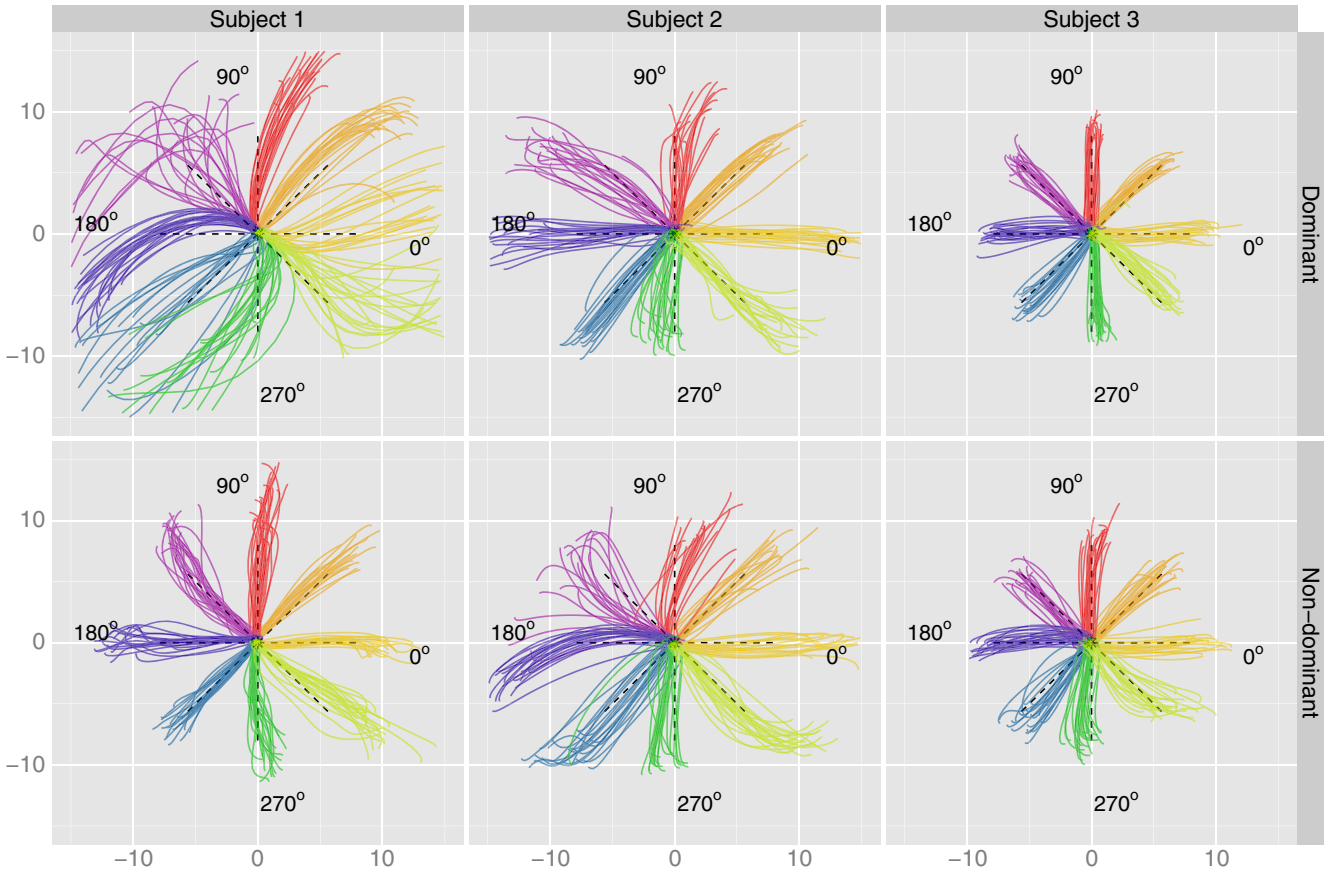

Figure 1. Observed reaching motions for three subjects. The top row shows the dominant hand, and the bottom row shows the non-dominant hand of three subjects. The left column is a subject with a contralesional dominant hand. The centre column is a subject with a contralesional non-dominant hand. The right column is a healthy control subject. Dashed lines are the straight paths to the eight targets terminating at the target location.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.106

The ISI's Journal for the Rapid Dissemination of Statistics Research

predictors x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$ and functional responses $y_i(t)$, $i = 1, \ldots, n$, $t \in \mathcal{T}$, where $\mathcal{T}$ is some compact finite interval in $\mathbb{R}$; the linear function-on-scalar regression model is

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{p} x_{ij}\beta_j(t) + \epsilon_i(t), \; i = 1, \ldots, n, \; t \in \mathcal{T} \tag{1}$$

where $\beta_j(\cdot), j = 0, \ldots, p$ are the $p + 1$ coefficient functions and $\epsilon_i(\cdot) \sim (0, \Sigma)$ is the error function drawn from a continuous stochastic process with expectation zero and covariance function $\Sigma(s, t) = \text{cov}(\epsilon_i(s), \epsilon_i(t))$, $s, t \in \mathcal{T}$.

A common model fitting framework for function-on-scalar regression is outlined by Chapter 13 of Ramsay & Silverman (2005), in which the coefficient functions $\beta_j(\cdot)$ are expanded using some set of basis functions and basis coefficients are estimated using ordinary least squares. The imposition of quadratic roughness penalties to enforce smoothness of the estimated coefficient functions is also common. Reiss et al. (2010) developed a fast automatic method for choosing tuning parameters in this model and accounted for correlated errors using generalized least squares. Goldsmith & Kitago (2016) develop a Bayesian approach that jointly models coefficient functions and the covariance structure and applied their methods to the stroke kinematics dataset considered here.

When p is large, many scalar predictors may have no effect on the functional response, and the corresponding coefficient functions would equal to zero over all time points. In order to accurately identify the important predictors, we apply variable selection techniques when estimating the coefficient functions in model (1). Because the coefficient functions are expanded using basis functions, the shape of each coefficient function is determined by a distinct group of basis coefficients. We therefore apply variable selection at the group level to include or exclude the vector of basis coefficients. The group lasso, proposed by Yuan & Lin (2006), is an extension of the classic lasso (Tibshirani, 1994) to the problem of selecting grouped variables. The lasso is known to induce biases in the included variables, so two alternative penalties, the smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001) and the minimax concave penalty (MCP) (Zhang, 2010), were proposed. These achieve consistency and asymptotic unbiasedness and have been extended to grouped variable selection problem (Wang et al., 2007; Breheny & Huang, 2013).

Few approaches that consider variable selection in the context of functional regression models have been proposed in current literature. Wang et al. (2007) developed a penalized estimation procedure using group SCAD for variable selection in function-on-scalar regression assuming errors $\epsilon_i(\cdot)$ that are uncorrelated over their domain; this assumption is clearly violated in practice. Barber et al. (2015) presented function-on-scalar LASSO (FS-LASSO), a framework which extends the group LASSO to function-on-scalar regression; theory is developed for cases in which predictors are observed over dense or sparse grids. However, the bias for non-zero coefficients introduced by LASSO was not addressed, and the method does not account for correlation among residual curves. Gertheiss et al. (2013) proposed a variable selection procedure for generalized scalar-on-function linear regression models, in which the predictors are in the form of functions and responses are scalar; although they also consider regression models for functional data, the structure of their models is very different from the one considered here.

The main contribution of this paper is a method for variable selection in function-on-scalar regression that accounts for residual correction using tools from generalized least squares. We develop theory for this method and demonstrate its effectiveness in simulations that mimic our real-data application; direct comparisons with the method of Wang et al. (2007) and Barber et al. (2015) indicate superior performance of our proposed method for variable selection and prediction.

The rest of the article is organized as follows. In Section 2, we describe an estimation procedure for function-on-scalar regression models with errors that are uncorrelated over t using grouped variable selection methods. We then introduce our methods for the estimation of function-on-scalar regression models with correlated errors, including the development of an iterative method that refines the estimation of the error covariance and the variable selection. Sim-

(wileyonlinelibrary.com) DOI: 10.1002/sta4.106

ulations that resemble our motivating data examine and compare the numerical performance of competing methods in Section 3. An application of our method to the reaching motion data is given in Section 4. Finally, we present concluding remarks in Section 5. Our method is implemented in the user-friendly fosr.vs() function in the *refund* package (Ciprian et al., 2014), available on CRAN, and code for our simulations is included in the supplementary material.

2 Methodology

2.1 Estimation for models with i.i.d. errors

Suppose $\{\phi_1(\cdot), \ldots, \phi_K(\cdot)\}$ is a set of pre-specified basis functions. The coefficient functions $\beta_j(\cdot)$, $j = 0, \ldots, p$ can be expanded as

$$\beta_j(\cdot) = \sum_{k=1}^{K} b_{jk}\phi_k(\cdot). \tag{2}$$

Hence, model (1) is expressed as

$$y_i(t) = \sum_{k=1}^{K} b_{0k}\phi_k(t) + \sum_{j=1}^{p} x_{ij}\left(\sum_{k=1}^{K} b_{jk}\phi_k(t)\right) + \epsilon_i(t). \tag{3}$$

The problem is thereby reduced to estimating the basis coefficients $\{b_{jk}\}_{j=0,\ldots,p;\, k=1,\ldots,K}$.

In practice, functions are observed on a discrete grid. For simplicity, we assume that the grid, denoted $\{t_1, \ldots, t_D\}$, is shared across subjects. Let $\boldsymbol{Y}$ be the $n \times D$ matrix whose rows are vector-valued functional responses; Φ be the $D \times K$ matrix whose columns correspond to the K basis functions evaluated at $\{t_1, \ldots, t_D\}$; and $\boldsymbol{B}$ be the $(p+1) \times K$ matrix with the jth row being the vector of basis coefficients for $\beta_j(\cdot)$. Then model (3) can be expressed as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}\Phi^T + \boldsymbol{E} \tag{4}$$

where $\boldsymbol{X}$ is the $n \times (p+1)$ design matrix and $\boldsymbol{E}$ is the $n \times D$ matrix containing vector-valued error functions.

Model (4) can be posed as a standard linear model in the following way. Let $\text{vec}(\boldsymbol{Y}^T)$ be the vector formed by concatenating the rows of $\boldsymbol{Y}$, and note that $\text{vec}\left((\boldsymbol{X}\boldsymbol{B}\Phi^T)^T\right) = (\boldsymbol{X} \otimes \Phi)\text{vec}(\boldsymbol{B}^T)$ where $\otimes$ represents the Kronecker product of two matrices. Then

$$\text{vec}\left(\boldsymbol{Y}^T\right) = (\boldsymbol{X} \otimes \Phi)\text{vec}\left(\boldsymbol{B}^T\right) + \text{vec}\left(\boldsymbol{E}^T\right) \tag{5}$$

and $\text{vec}(\boldsymbol{B}^T)$ can be estimated using least squares. An estimate of $\hat{\boldsymbol{B}}$ is obtained by rearranging $\text{vec}\left(\hat{\boldsymbol{B}}^T\right)$.

To accurately identify the zero coefficient functions, we apply variable selection techniques when estimating $\text{vec}(\boldsymbol{B}^T)$ in model (5). Let $\boldsymbol{B}_j$ be the vector of coefficients associated with the jth coefficient function $\beta_j(\cdot)$, specifically the jth row of $\boldsymbol{B}$. Note that the "zeroth" row of $\boldsymbol{B}$ corresponds to the intercept function $\beta_0(t)$, which we do not penalize. Setting the entire $\beta_j(\cdot)$ function to 0 is equivalent to setting all the entries of $\boldsymbol{B}_j$ to zero. Therefore, we apply variable selection techniques at the group level.

Variable selection can be achieved by penalizing the estimates of the coefficients. The general form of a group penalty is $\sum_{j=1}^{p} p_{\lambda,\gamma}(||\boldsymbol{B}_j||)$, where $p_{\lambda,\gamma}(\cdot)$ is the penalty function for the specific method and λ and γ are the tuning parameters. Therefore, the penalized estimator is obtained by minimizing

$$\frac{1}{2}\left\|\text{vec}(\boldsymbol{Y}^T) - (\boldsymbol{X} \otimes \Phi)\text{vec}(\boldsymbol{B}^T)\right\|^2 + nD\sum_{j=1}^{p} p_{\lambda,\gamma}(||\boldsymbol{B}_j||). \tag{6}$$

(wileyonlinelibrary.com) DOI: 10.1002/sta4.106

The ISI's Journal for the Rapid Dissemination of Statistics Research

We use group MCP to perform variable selection; the penalty has the form

$$p_{\text{mcp}}(||\boldsymbol{B}_j||) = \begin{cases} \lambda||\boldsymbol{B}_j|| - \frac{||\boldsymbol{B}_j||^2}{2\gamma} & \text{if } ||\boldsymbol{B}_j|| \leq \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2 & \text{if } ||\boldsymbol{B}_j|| > \gamma\lambda \end{cases}$$

where λ and γ are the tuning parameters. When $||\boldsymbol{B}_j||$ is small, the MCP penalty behaves exactly as lasso, but as $||\boldsymbol{B}_j||$ increases, the amount of penalization is reduced until there is no penalization at all, thereby avoiding bias in the estimate of large coefficients.

In terms of tuning parameter selection, γ is set to be 3 as recommended in Zhang (2010) and λ is chosen by cross-validation. Another parameter to be determined is K, the number of basis functions used in the expansion of the coefficient functions. In the following implementations of our method in Section 3, a cubic B-spline basis with 10 basis functions was used. However, because we do not explicitly penalize the roughness of the estimated coefficient functions, the exact choice of K will vary from application to application and should be chosen with care. In Section 4, we use cross-validation to choose K.

2.2 Estimation for models with correlated errors

The estimation framework discussed in Section 2.1 assumes that errors are independent and identically distributed over the entire domain and is similar to the framework of Wang et al. (2007). In most cases, however, within-function errors are correlated. Let Σ denote the $D \times D$ covariance matrix for discretely observed data. For estimation of the model (4) with correlated errors, we use techniques from generalized least squares. If Σ is known, one can "pre-white" both sides of (4) with the lower triangular matrix $\boldsymbol{L}$ obtained by Cholesky decomposition of Σ, that is, $\Sigma = \boldsymbol{L}\boldsymbol{L}^T$, to construct a new model

$$\boldsymbol{Y}^* = \boldsymbol{X}\boldsymbol{B}\Phi^{*T} + \boldsymbol{E}^* \tag{7}$$

where $\boldsymbol{Y}^* = \boldsymbol{Y}(\boldsymbol{L}^{-1})^T$, $\Phi^* = \boldsymbol{L}^{-1}\Phi$ and the error $\boldsymbol{E}^* = \boldsymbol{E}(\boldsymbol{L}^{-1})^T$ is independent. Similarly, parameters in model (7) can be estimated by minimizing

$$\frac{1}{2}\left\|\text{vec}(\boldsymbol{Y}^{*T}) - (\boldsymbol{X} \otimes \Phi^*)\text{vec}(\boldsymbol{B}^T)\right\|^2 + nD\sum_{j=1}^{p} p_{\lambda,\gamma}(||\boldsymbol{B}_j||). \tag{8}$$

For a given Σ, the minimizer of (8) can be obtained using existing software by pre-whitening as described; our implementation is publicly available and uses the *grpreg* function in the *grpreg* package (Breheny & Huang, 2013).

The covariance matrix Σ is unknown in practice, and it is necessary to obtain an estimate $\hat{\Sigma}$ of Σ and to use this estimate to pre-whiten data. To obtain this estimate, we first fit model (5) using ordinary least squares under the assumption of independence; this provides an unbiased estimate $\hat{\boldsymbol{B}}$ of the coefficient matrix $\boldsymbol{B}$. From this model fit, we obtain the estimated residual matrix $\hat{\boldsymbol{E}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}\Phi^T$. Using $\hat{\boldsymbol{E}}$, we consider two approaches for estimating Σ. The first, which we refer as the raw estimate, is constructed using a method-of-moments approach based on the residual matrix. The second approach uses functional principal component analysis (Yao et al., 2005). Here, the off-diagonal elements of the raw covariance are smoothed and an eigendecomposition of the resulting matrix is obtained. Our estimate is

$$\hat{\Sigma} = \sum_{l=1}^{L} \hat{\nu}_l \hat{\boldsymbol{\psi}}_l \hat{\boldsymbol{\psi}}_l^T + \hat{\sigma}^2 \boldsymbol{I} \tag{9}$$

where $\hat{\boldsymbol{\psi}}_1, \ldots, \hat{\boldsymbol{\psi}}_L$ are the estimated eigenfunctions over the grid $\{t_1, \ldots, t_D\}$, $\hat{\nu}_l$, $l = 1, 2, \ldots, L$ are the corresponding eigenvalues, $\hat{\sigma}^2$ is the estimated measurement error variance and I is the identity matrix. The truncation level L is

(wileyonlinelibrary.com) DOI: 10.1002/sta4.106

determined by the cumulative proportion of variability explained by eigenfunctions. This approach separates Σ into a smooth covariance over the observed grid and an additional uncorrelated measurement error process. Although we focus on these methods for estimating Σ, others that provide consistent estimators can be substituted.

2.3 Oracle properties of generalized group minimax concave penalty estimator

We now discuss the theoretical properties of the method described in Section 2.2. Without loss of generality, we assume $\beta_0(t) = 0$, $\forall\, t \in \mathcal{T}$. Hence, $\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$ is an $n \times p$ matrix with the ith row being $\boldsymbol{x}_i^T = (x_{i1}, \ldots, x_{ip})$. We also assume the true coefficient functions $\beta_j(t)$ are in the space spanned by the set of basis functions Φ. Additionally, we assume the first s groups of coefficients, $\boldsymbol{B}_+ = (\boldsymbol{B}_1^T, \ldots, \boldsymbol{B}_s^T)^T$, are non-zero, and the remaining $p - s$ groups of coefficients, $\boldsymbol{B}_0 = (\boldsymbol{B}_{s+1}^T, \ldots, \boldsymbol{B}_p^T)^T$, are zero. Let $\boldsymbol{x}_{i+}$ denote the vector associated with $\boldsymbol{B}_+$ and $\boldsymbol{x}_{i0}$ denote the one associated with $\boldsymbol{B}_0$. Therefore, we have $\boldsymbol{B} = \left(\boldsymbol{B}_+^T, \boldsymbol{B}_0^T\right)^T$ and $\boldsymbol{x}_i^T = (\boldsymbol{x}_{i+}^T, \boldsymbol{x}_{i0}^T)$, where $\boldsymbol{x}_{i+}^T = (x_{i1}, \ldots, x_{is})$ and $\boldsymbol{x}_{i0}^T = (x_{i(s+1)}, \ldots, x_{ip})$. We further assume that the tuning parameter γ of the penalty is fixed. The additional conditions required for the theorems are

1. $\lim_{n\to\infty} \frac{1}{nD} \sum_{i=1}^n (\boldsymbol{x}_i^T \otimes \Phi)^T (\boldsymbol{x}_i^T \otimes \Phi)$ is a positive definite matrix;
2. $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$;
3. There exists an estimate $\hat{\Sigma}$ of Σ for which each element is $\sqrt{n}$-consistent;
4. Σ is non-singular.

Then we have the following results.

Theorem 1 (Estimation consistency)
Under Assumptions 1–4, there exists a local minimizer $\hat{\boldsymbol{B}}$ of

$$Q(\boldsymbol{B}) = \frac{1}{2}\left[\mathrm{vec}(\boldsymbol{Y}^T) - (\boldsymbol{X} \otimes \Phi)\mathrm{vec}(\boldsymbol{B}^T)\right]^T (\boldsymbol{I}_n \otimes \hat{\Sigma})^{-1} \left[\mathrm{vec}(\boldsymbol{Y}^T) - (\boldsymbol{X} \otimes \Phi)\mathrm{vec}(\boldsymbol{B}^T)\right] + nD \sum_{j=1}^{p} p_{\lambda_n,\gamma}(||\boldsymbol{B}_j||)$$

such that $\left\|\mathrm{vec}(\hat{\boldsymbol{B}}^T) - \mathrm{vec}(\boldsymbol{B}^T)\right\| = O_p(n^{-1/2})$.

Theorem 2 (Oracle property)
Under Assumptions 1–4, the $\sqrt{n}$-consistent local minimizer $\hat{\boldsymbol{B}} = \left(\hat{\boldsymbol{B}}_+^T, \hat{\boldsymbol{B}}_0^T\right)^T$ satisfies

(1) Sparsity: $\hat{\boldsymbol{B}}_0 = \boldsymbol{0}$, with probability tending to 1;
(2) Asymptotic normality:

$$\sqrt{n}\left(\mathrm{vec}\left(\hat{\boldsymbol{B}}_+^T\right) - \mathrm{vec}\left(\boldsymbol{B}_+^T\right)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\boldsymbol{0}, \left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_{i+}^T \otimes \Phi\right)^T \Sigma^{-1} \left(\boldsymbol{x}_{i+}^T \otimes \Phi\right)\right)^{-1}\right).$$

The proof of these theorems is provided in Online Supplement.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.106

The ISI's Journal for the Rapid Dissemination of Statistics Research

2.4 Iterative algorithm for models with correlated errors

The method described in Section 2.2 uses ordinary least squares to estimate basis coefficients and obtains an estimate $\hat{\Sigma}$ of the covariance Σ; this estimate is then used to pre-whiten the data prior to the application of variable selection techniques. However, re-estimating the covariance after variable selection may give a refined estimate which can, in turn, be used to pre-whiten the data. This intuition suggests an iterative algorithm:

1. Fit a model using ordinary least squares to obtain an initial estimate $\hat{\boldsymbol{B}}^{(0)}$;
2. Compute residuals and obtain an estimate $\hat{\Sigma}^{(0)}$ of Σ;
3. For $k > 0$, iterate the following steps until convergence:
   (a) Pre-whiten using the covariance $\hat{\Sigma}^{(k-1)}$;
   (b) Minimize (8) to obtain $\hat{\boldsymbol{B}}^{(k)}$;
   (c) Use $\hat{\boldsymbol{B}}^{(k)}$ to construct fitted values and residual curves, and use these to construct $\hat{\Sigma}^{(k)}$.

Various criteria of convergence can be used to monitor convergence of this iterative algorithm; one possible criterion is $\left|\left|\hat{\boldsymbol{B}}^{(k+1)} - \hat{\boldsymbol{B}}^{(k)}\right|\right|^2 < \delta$, which we use in our implementations. This iterative method will be compared with the one-step approach of Section 2.2 in simulations.

3 Simulation

We conducted simulation studies to examine the properties of the proposed approach. Specifically, we constructed 500 training samples, each consisting of 100 random curves, and 1 test sample containing 1000 random curves. All curves are generated from the model

$$y_i(t) = \sum_{j=1}^{20} x_{ij}\beta_j(t) + \epsilon_i(t)$$

where $x_{ij} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 10)$, $\beta_1(t)$, $\beta_2(t)$, $\beta_3(t)$ are non-zero functions, and the remaining coefficient functions are zero. All functions are observed on an equally spaced grid of length 25. Errors $\epsilon_i(t_d)$ are generated from a multivariate Gaussian distribution with mean zero and covariance $\Sigma = \boldsymbol{G} + \boldsymbol{I}$, where $\boldsymbol{G}$ is the error covariance and $\boldsymbol{I}$ is the identity matrix. the non-zero coefficient functions $\beta_1(t)$, $\beta_2(t)$ and $\beta_3(t)$ are derived from the motivating data in the following way. Focusing on y position curves for reaching motions made to the target at 0 degrees, we estimated motions made by healthy controls, moderately affected stroke patients and severely affected patients (stroke severity was defined by thresholding the Fugl-Meyer score). These estimated motions were the non-zero coefficients and are shown in the middle panel of Figure 2. The error covariance G was constructed using a functional principal component analysis (FPCA) decomposition of residual curves after subtracting the group-specific means.

Four implementations of our proposed method are considered: one-step approaches as described in Section 2.2 using raw and FPCA-based covariance matrix estimates, and iterative approaches as described in Section 2.4 using raw and FPCA-based covariance matrix estimates. For the FPCA-based covariance matrix estimate, we used two different values, 0.5 and 0.99, as the cumulative proportion of variance explained (PVE) threshold to determine L. For comparison, we include an approach that pre-whitens using true covariance matrix, as well as ordinary least squares, a variational Bayes method that includes a smoothness penalty (Goldsmith & Kitago, 2016), the FS-LASSO method that uses group LASSO but does not account for residual correlation or biases due to the LASSO penalty and a group MCP method that assumes uncorrelated error curves, analogously to Wang et al. (2007).

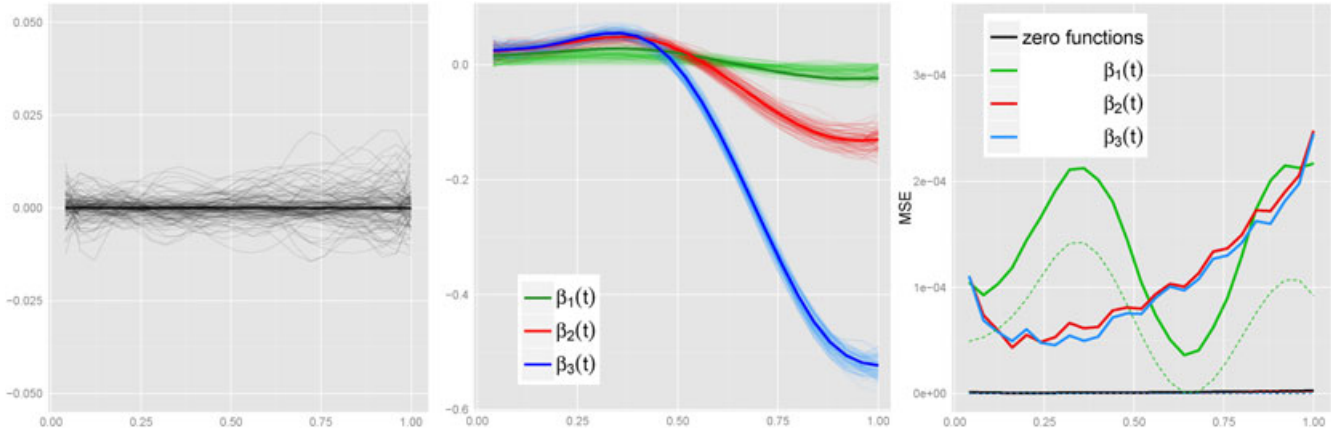

Figure 2. Estimates of zero functions (left) and non-zero functions (middle) obtained using the iterative approach with FPCA-based covariance matrix estimate using PVE = 0.99 across all simulated datasets. The true functions are overlaid (bold curves). The right panel shows both MSE (solid) and squared bias (dashed) as functions of time for all the coefficient functions.

Table I. True positive (TP) and true negative (TN) rates of estimated coefficient functions, where FN is false negative and FP is false positive. They are estimated across all the training samples.

	$\frac{TN}{FP+TN}$	$\frac{TP}{TP+FN}(\beta_1)$	$\frac{TP}{TP+FN}(\beta_2)$	$\frac{TP}{TP+FN}(\beta_3)$
FS-LASSO	0.953	0.850	1.000	1.000
MCP assuming independent errors	0.567	1.000	1.000	1.000
One-step with raw matrix	0.370	1.000	1.000	1.000
Iterative with raw matrix	0.813	0.962	1.000	1.000
One-step with FPCA-based matrix (PVE = 0.5)	0.755	0.996	1.000	1.000
Iterative with FPCA-based matrix (PVE = 0.5)	0.779	0.996	1.000	1.000
One-step with FPCA-based matrix (PVE = 0.99)	0.863	0.986	1.000	1.000
Iterative with FPCA-based matrix (PVE = 0.99)	0.915	0.956	1.000	1.000
Pre-whiten with true Σ	0.913	0.964	1.000	1.000

Table I reports the true positive (TP) and true negative (TN) rates of the estimates of both zero and non-zero coefficient functions. We define functions estimated to be non-zero as "positive" while functions estimated to be zero as "negative". Our iterative approach using an FPCA-based covariance matrix estimate with PVE = 0.99 outperforms most competing approaches in terms of correctly identifying the zero functions; its performance is comparable with the approach that uses the true covariance matrix. The approaches using PVE = 0.5 perform less well because the estimate of the covariance matrix omits important structure. Our proposed methods substantially outperform the method that assumes uncorrelated errors in accurately identifying zero functions. FS-LASSO has the highest true negative rate but the lowest true positive rate for $\beta_1(t)$, potentially indicating a tendency to over-shrink coefficients to zero. All methods are able to identify $\beta_2(t)$ and $\beta_3(t)$ as non-zero.

Estimates of zero and non-zero coefficient functions obtained using the iterative algorithm with FPCA-based covariance matrix estimate using PVE = 0.99 are shown in the left and middle panels of Figure 2, respectively. Because their coefficients are relatively large, the estimate of $\beta_2(\cdot)$ and $\beta_3(\cdot)$ are approximately unbiased owing to the structure of the

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.106

The ISI's Journal for the Rapid Dissemination of Statistics Research

penalty. For $\beta_1(\cdot)$, coefficients are shrunk toward and sometimes set equal to zero. We show the mean squared error MSE $= E\left(\beta_j(t) - \hat{\beta}_j(t)\right)^2$ and squared bias $E\left(\beta_j(t) - \bar{\hat{\beta}}_j(t)\right)^2$ as functions of t in the right panel of Figure 2, where $\bar{\hat{\beta}}_j(t)$ is the average curve across all the simulation datasets. For $\beta_1(\cdot)$, both the MSE and squared bias curves present a sinusoidal shape, which is driven by the sinusoidal shape of the coefficient function itself and by the shrinkage to zero. There is an increasing trend in general as t increases for the MSE of $\beta_2(\cdot)$ and $\beta_3(\cdot)$, which is mostly caused by the increased variability of curves at the end of the distribution as the biases are relatively small. This plot further emphasizes the lack of bias for large coefficients stemming from the use of the group MCP penalty, especially in the case of $\beta_3(\cdot)$.

The left and middle columns of Figure 3 display the root mean integrated squared error (RMISE), $\sqrt{\int_0^1 \left(\beta_j(t) - \hat{\beta}_j(t)\right)^2 dt}$ for zero and non-zero functions, respectively; in the top row, the FPCA-based covariance estimate is based on PVE = 0.99 and in the bottom row based on PVE = 0.5. The iterative approach with FPCA-based covariance matrix estimate compares favourably to other approaches, reinforcing the results from Table I. Indeed, the

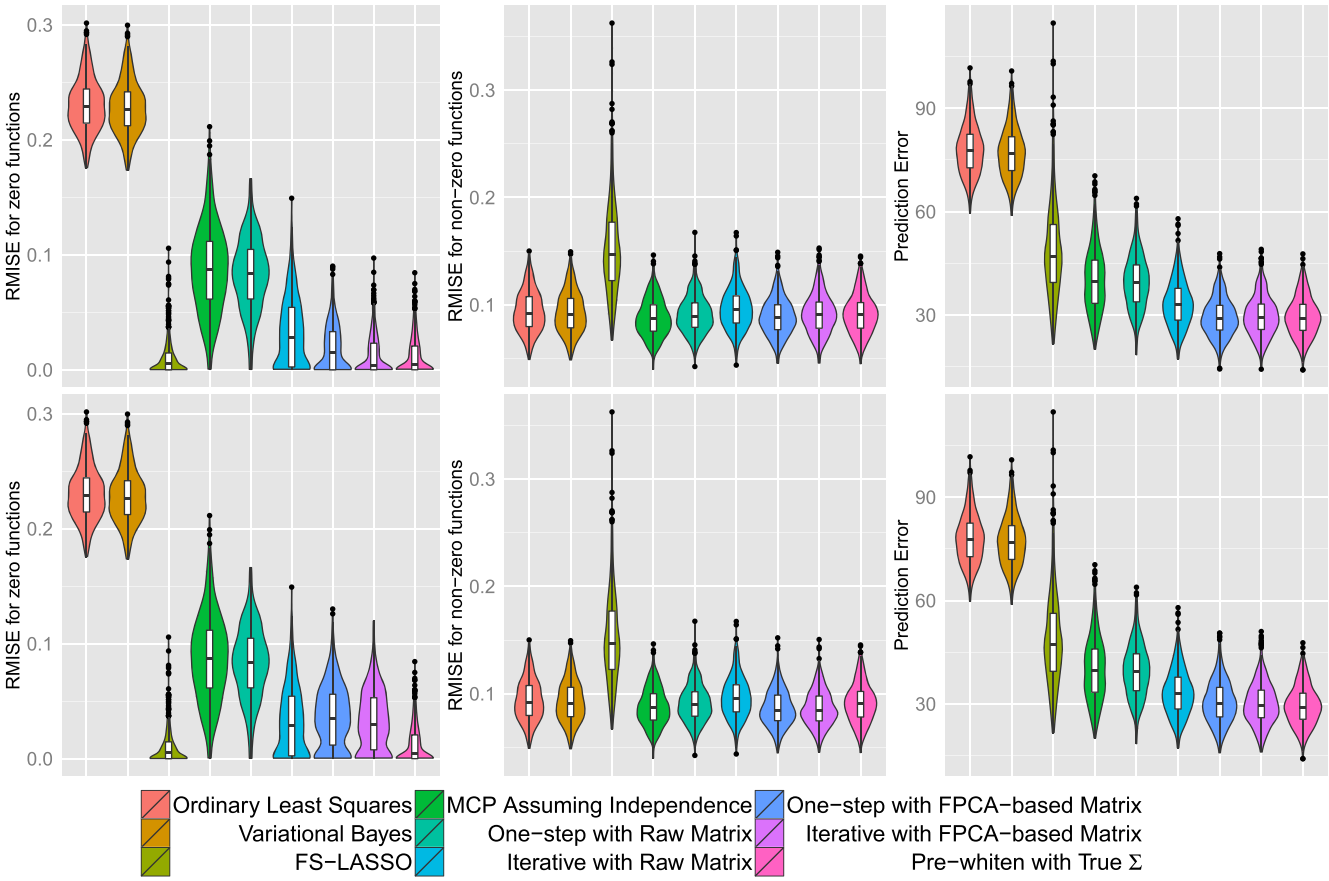

Figure 3. The top row shows the comparison among the algorithms when PVE = 0.99, while the second row shows the comparison when PVE = 0.5. The three columns show RMISE for zero functions (left) and non-zero functions (middle) and prediction error (right).

RMISE of our iterative method is comparable with pre-whitening using the true covariance for both zero and non-zero functions. Although FS-LASSO is comparable for zero functions, it has substantially higher RMISE for non-zero functions. Prediction errors on the test sample are shown in the right panel of Figure 3. These errors reflect a combination of RMISEs for zero and non-zero functions and display similar patterns: our proposed methods, in particular when using the FPCA-based estimate of the covariance, have excellent numerical performance. Although there is a slight decline in performance when PVE = 0.5, the proposed method still outperforms ordinary least square, FS-LASSO and the method that assumes uncorrelated errors.

Additional simulations that generate uncorrelated errors are presented in detail in Online Supplement. In this case, there is no noticeable disadvantage to using our proposed approach, which outperforms competing methods in prediction error.

4 Application

We now apply our iterative algorithm using the FPCA-based covariance matrix estimate described in Section 2.4 to our motivating dataset. The *X* and *Y* position functions are the outcomes of interest, and potential predictors include the Fugl-Meyer score, whether the hand was dominant or non-dominant, whether the hand was contralesional or ipsilesional, target direction (as a categorical predictor) and the interactions between these variables. Goldsmith & Kitago (2016) modelled *X* and *Y* position curves as a bivariate process; the results of that analysis indicated relatively low correlation between residuals in the *X* and *Y* directions. We therefore model *X* and *Y* position curves separately, using the same models and steps.

First, we perform a cross-validation analysis to evaluate the algorithm in terms of prediction error on the motivating data. Training and test sets are generated in the following way. For each subject and each hand, we randomly select one motion to each of the eight target directions. These motions are partitioned so that four are in the training set and four are in the test set. Previous work on this dataset (Goldsmith & Kitago, 2016) indicates little or no correlation between motions to different targets made by the same subject, and so our training and test sets are approximately independent even though they contain data from the same subjects. This procedure results in 452 curves in the training set and 452 curves in the test set; an example is shown in Figure 4.

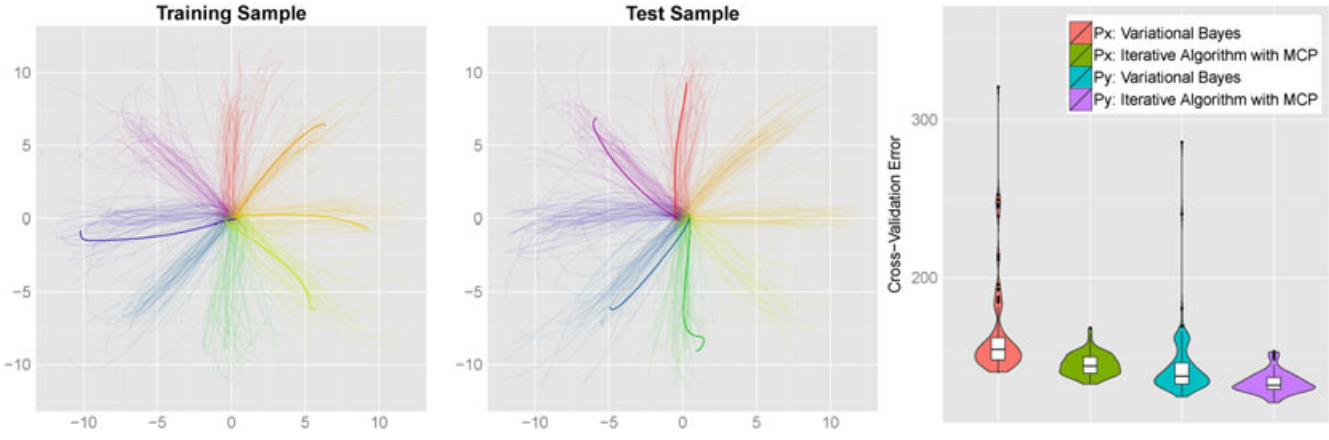

Figure 4. One training sample (left) and one test sample (middle) generated from the planar reaching data. Highlighted curves are from one subject and show how each subject contributes to the training and test sets. Violin plots (right) of cross-validation errors using the variational Bayes approach and iterative algorithm.

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.106

The ISI's Journal for the Rapid Dissemination of Statistics Research

A function-on-scalar regression model is then constructed on the training sample, and prediction errors are obtained for the test sample. Four predictors of interest, the target direction (a categorical variable with eight levels), Fugl-Meyer score (a continuous variable), hand used (dominant/non-dominant) and arm affectedness (contralesional/ipsilesional), are considered in these models. In addition to main effects, all the possible interactions are included to maximize flexibility and scientific interpretation. Thus, the model has 64 coefficient functions to estimate. Rather than the typical design that assigns a reference level for each categorical predictor, a constraint is imposed to the construction of design matrix so that target-specific interpretations are available. This design matrix is equivalent to building the following model for each target:

$$\begin{aligned} y(t) = {} & \beta_0(t) + \beta_1(t) * \text{Ips.Non.} + \beta_2(t) * \text{Con.Dom.} + \beta_3(t) * \text{Con.Non.} + \beta_4(t) * \text{Fugl-Meyer} \\ & + \beta_5(t) * \text{Fugl-Meyer} * \text{Ips.Non.} + \beta_6(t) * \text{Fugl-Meyer} * \text{Con.Dom.} \\ & + \beta_7(t) * \text{Fugl-Meyer} * \text{Con.Non.} + \epsilon(t) \end{aligned} \tag{10}$$

where we use the ipsilesional (unaffected) dominant hand of a healthy control as the reference $\beta_0(t)$. Coefficients $\beta_1(t)$, $\beta_2(t)$ and $\beta_3(t)$ compare ipsilesional non-dominant, contralesional dominant and contralesional non-dominant to the reference, respectively. The effect of increasing motor impairment in the ipsilesional dominant arm is estimated by $\beta_5(t)$, while differences in the effect of increasing motor impairment comparing other groups to baseline are given by $\beta_6(t)$, $\beta_7(t)$ and $\beta_8(t)$.

The complete procedure described earlier, consisting of generating training and test sets, fitting the full model to the training set, and producing predictions for the test set, is repeated 100 times. We fit the model using 5, 10, 15 and

Table II. Proportions of 64 coefficient functions being selected, obtained from models with *X* trajectories (top) and *Y* trajectories (bottom).

Target	Fugl-Meyer = 66				Δ Fugl-Meyer = −1			
direction	Ips.Dom.	Ips.Non.	Con.Dom.	Con.Non.	Ips.Dom.	Ips.Non.	Con.Dom.	Con.Non.
0°	1.00	0.21	0.24	0.35	0.41	0.38	0.58	0.34
45°	1.00	0.20	0.03	0.15	0.16	0.08	0.46	0.37
90°	0.31	0.65	0.31	0.33	0.30	0.17	0.23	0.60
135°	1.00	0.22	0.16	0.24	0.11	0.48	0.64	0.57
180°	1.00	0.35	0.13	0.40	0.18	0.38	0.48	0.35
225°	1.00	0.18	0.16	0.33	0.08	0.22	0.45	0.36
270°	0.83	0.37	0.19	0.34	0.14	0.35	0.52	0.33
315°	1.00	0.41	0.20	0.33	0.28	0.35	0.57	0.67
Target	Fugl-Meyer = 66				Δ Fugl-Meyer = −1			
direction	Ips.Dom.	Ips.Non.	Con.Dom.	Con.Non.	Ips.Dom.	Ips.Non.	Con.Dom.	Con.Non.
0°	0.34	0.10	0.14	0.11	0.13	0.18	0.66	0.20
45°	1.00	0.02	0.02	0.05	0.01	0.20	0.36	0.12
90°	1.00	0.07	0.16	0.15	0.08	0.21	0.17	0.43
135°	1.00	0.05	0.22	0.29	0.05	0.40	0.79	0.43
180°	0.40	0.13	0.11	0.31	0.15	0.25	0.79	0.27
225°	1.00	0.05	0.03	0.11	0.01	0.12	0.33	0.21
270°	1.00	0.03	0.15	0.14	0.06	0.24	0.13	0.22
315°	1.00	0.03	0.22	0.18	0.06	0.36	0.52	0.54

20 basis functions, and found that $K = 15$ gave the smallest cross-validated prediction errors although $K = 10$ and $K = 20$ were both very similar. The right panel of Figure 4 presents the prediction errors obtained using our iterative algorithm with FPCA-based covariance matrix estimate; we compare to the variational Bayes approach (without variable selection but with a standard second-derivative penalty). Our iterative algorithm decreases mean prediction error by around 10% (*X* direction: 163.8 vs. 144.8; *Y* direction: 143.6 vs. 132.9) compared with the variational Bayes approach. In addition, the iterative algorithm seems to be more stable than the variational Bayes approach as it has fewer outliers and lower median prediction error.

We next conduct our analysis without splitting data into training and test sets. The function-on-scalar regression model is estimated using one motion for each subject and hand to each target with motions drawn randomly for each target and hand. We repeat this analysis 100 times, and Table II presents the proportion of times selected by the algorithm for each of the 64 coefficient functions. Each row of Table II corresponds to coefficients $\beta_0(t), \beta_1(t), \ldots, \beta_7(t)$ in model (10) for a specific target. For instance, the value 0.24 in the third entry of the first row indicates that in 24 of the 100 datasets, there was an estimated difference between contralesional and ipsilesional dominant hands when reaching to the target at $0°$.

Large numbers in the table suggest consistent non-zero effects or differences in effect across datasets. Targets at $90°$ and $270°$ may have zero effects in the *X* trajectories, because for those targets, the *X* position is roughly constant over time. The same is true for targets at $0°$ and $180°$ for the *Y* trajectories. The results in Table II indicate relatively few differences between ipsilesional and contralesional dominant arms for very mild strokes (Fugl-Meyer = 66), and some differences between the non-dominant arms and the ipsilesional dominant arm. An effect of increasing stroke severity is relatively rarely found for the ipsilesional arms but, as expected, is much more frequently found for the

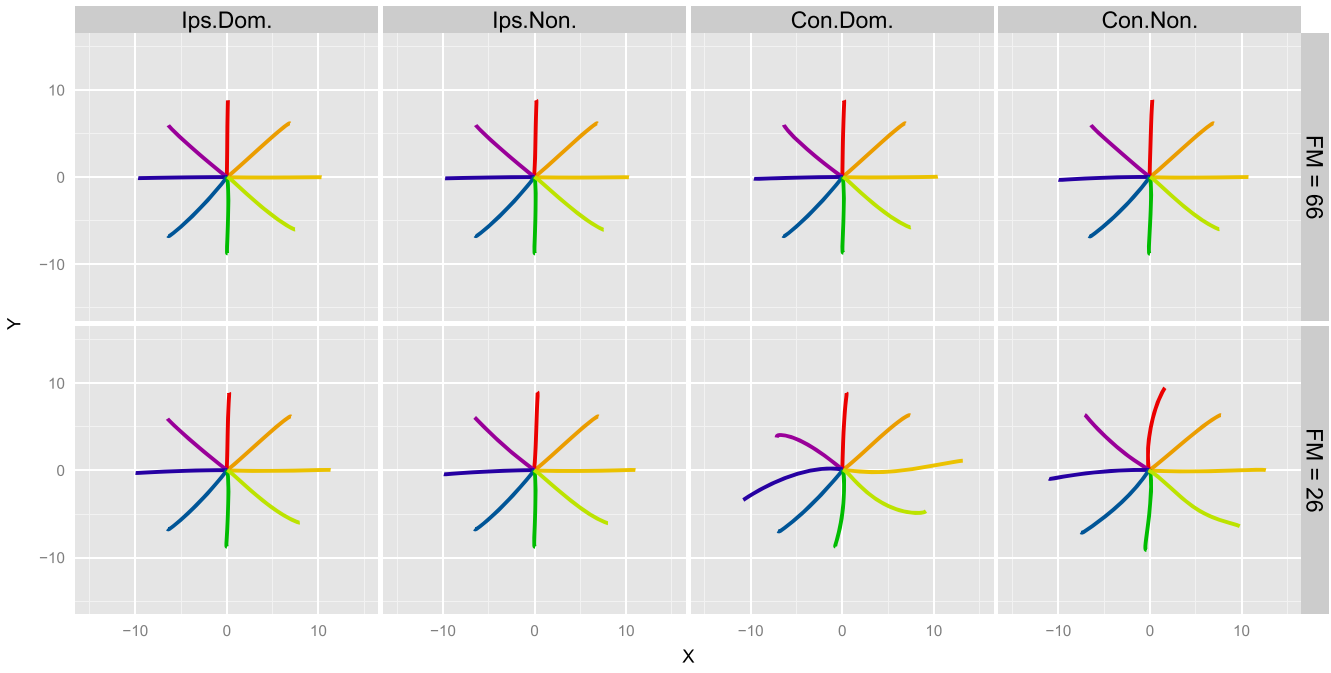

Figure 5. Predicted reaching motions for eight subjects with different combinations of Fugl-Meyer Score(66/26), hand used (dominant/non-dominant) and arm affectedness (contralesional/ipsilesional). Motions to different targets are distinguished by colours.

contralesional arms. The conclusions are further reinforced by Figure 5, where the predicted motions of subjects with different combinations of Fugl-Meyer Score (66/26), hand used (dominant/non-dominant) and arm affectedness (contralesional/ipsilesional) are presented.

5 Discussion

We proposed a model fitting framework that performs variable selection in the context of function-on-scalar regression allowing within-function correlation. This work was motivated by two-dimensional planar reaching data gathered to understand the mechanisms of motor deficit following stroke. We developed an iterative algorithm that alternatively estimates the coefficient functions and covariance structure. Our method relies on a reasonable estimate of the covariance structure, and in our simulations and application, we found that an estimation procedure based on FPCA works well. Results from the simulation studies demonstrate the effectiveness of our proposed method in identifying the true zero functions. Indeed, our proposed method has performance comparable with performing variable selection using the true covariance. The application to the motivating data indicates our proposed iterative algorithm makes a significant improvement in terms of decreasing prediction errors and identifying true zero functions.

Future extension of our methodology may take several directions. Quadratic roughness penalties are often applied to enforce smoothness of the coefficient functions in spline-based estimation frameworks. It would be worthwhile to incorporate an explicit roughness penalty in addition to the variable selection penalty to reduce sensitivity to the size of the basis expansion. Motivated by our application (in which repeated motions are made to each target by each subject), the development of methods that account for subject-specific and target-specific random effects is necessary.

References

Fugl-Meyer, AR, Jaasko, L, Leyman, I, Olsson, S & Steglind, S (1975), 'The post-stroke hemiplegic patient: a method for evaluation of physical performance', *Scandinavian Journal of Rehabilitation Medicine*, **7**(1), 13–31.

Ramsay, JO & Silverman, BW (2005), *Functional Data Analysis*, 2nd edn., Springer Series in Statistics, *Springer*, Hardcover.

Reiss, PT, Huang, L & Mennes, M (2010), 'Fast function-on-scalar regression with penalized basis expansions', *International Journal of Biostatistics*, **6**(1), article–28.

Goldsmith, J & Kitago, T (2016), 'Assessing systematic effects of stroke on motor control using hierarchical function-on-scalar regression', *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **65**(2), 215–236.

Yuan, M & Lin, Y (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society: Series B*, **68**(1), 49–67.

Tibshirani, R (1994), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**, 267–288.

Fan, J & Li, R (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association*, **96**(456), 1348–1360.

Zhang, CH (2010), 'Nearly unbiased variable selection under minimax concave penalty', *The Annals of Statistics*, **38**(2), 894–942.

Wang, L, Chen, G & Li, H (2007), 'Group SCAD regression analysis for microarray time course gene expression data', *Bioinformatics*, **23**(12), 1486–1494.

Breheny, P & Huang, J (2013), 'Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors', *Statistics and Computing*, **23**(6), 1–15.

Barber, RF, Reimherr, M & Schill, T,. (2015)*, The function-on-scalar LASSO with applications to longitudinal GWAS*, Technical Report.

Gertheiss, J, Maity, A & Staicu, AM (2013), 'Variable selection in generalized functional linear models', *Stat*, **2**(1), 86–101.

Ciprian, C, Philip, R, Jeff, G, Lei, H, Lan, H & Fabian, S (2014), *Refund: Regression with functional data*. Available from: http://CRAN.R-project.org/package=refund [Accessed on 24 February 2016], R package version 0.1-11.

Yao, F, Müller, HG & Wang, JL (2005), 'Functional data analysis for sparse longitudinal data', *Journal of the American Statistical Association*, **100**(470), 577–590. Available from: http://www.jstor.org/stable/27590579 [Accessed on 24 February 2016].

Zeng, L & Xie, J (2014), 'Group variable selection via SCAD-L2', *Statistics*, **48**(1), 49–66.

Peng, H & Lu, Y (2012), 'Model selection in linear mixed effect models', *Journal of Multivariate Analysis*, **109**(0), 109 –129. Available from: http://www.sciencedirect.com/science/article/pii/S0047259X12000395 [Accessed on 24 February 2016].

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site.