VARIABLE SELECTION FOR HISTORICAL FUNCTIONAL LINEAR MODEL

Matsui, Hidetoshi
Faculty of Data Science, Shiga University

https://doi.org/10.5109/4151124

出版情報：Bulletin of informatics and cybernetics. 53 (1), pp.1-19, 2021. 統計科学研究会
バージョン：
権利関係：

VARIABLE SELECTION FOR HISTORICAL FUNCTIONAL LINEAR MODEL

by

Hidetoshi Matsui

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.53, No. 1

FUKUOKA, JAPAN
2021

Bulletin of Informatics and Cybernetics, Vol. 53, No. 1, 2021

VARIABLE SELECTION FOR HISTORICAL FUNCTIONAL LINEAR MODEL

By

Hidetoshi MATSUI[*]

Abstract

We consider a variable selection problem for functional linear models where both multiple predictors and a response are functions. We assume that these variables are given as functions of time and then construct the historical functional linear model which takes the dependence of time between multiple predictors and a response into consideration. Unknown parameters included in the model are estimated by the maximum penalized likelihood method with the L_1-type penalty. We can simultaneously estimate and select variables given as functions owing to the sparsity penalty. A regularization parameter involved in the regularization method is decided by a model selection criterion. The effectiveness of the proposed method is investigated by simulation studies and real data analysis.

Key Words and Phrases: Functional data analysis, Longitudinal data; Model selection; Sparse regularization.

1. Introduction

Functional data analysis (FDA) has received considerable attention in several fields such as meteorology, ergonomics and medicine, and there have been many studies on both its theory and applications (see, e.g., Ramsay and Silverman, 2002, 2005; Kokoszka and Reimherr, 2017). The basic concept behind FDA is to represent repeated measurement data for individuals as smooth functions and then treat the individuals as if they themselves were the observed data.

There has been much work on regression modelings where both or either predictors and/or a response are given as functions. When the predictors are functions while the response is a scalar, such functional regression models have been discussed in several frameworks, such as the generalized linear models (James, 2002; Müller and Stadtmüller, 2005), the additive models (Müller and Yao, 2008; Fan et al., 2015), and the adaptive models (James and Silverman, 2005). On the other hand, when both the predictors and the response are functions, the modeling strategy is further divided into two types. One is that the arguments of the functions are common (respectively denoted by $X(t)$ and $Y(t)$), and in the other case, they are functions of different arguments ($X(s)$ and $Y(t)$). The former case is so called the functional concurrent model and for this case the varying coefficient model (Hastie and Tibshirani, 1993) can be also applied to model the relationship between X and Y. In the latter case, the functional predictors affect the response over the domain and it is a more general model than the former case.

[*] Faculty of Data Science, Shiga University 1-1-1 Banba Hikone Shiga 522-8522 Japan. tel +81–749–27–1295 hmatsui@biwako.shiga-u.ac.jp

We consider the latter case throughout this paper. Ramsay and Dalzell (1991) first constructed the linear model for a functional predictor and a functional response as follows. Let $X(s)$ and $Y(t)$ be a predictor and a response, respectively, and functions in general Hilbert spaces with $s \in [0, S]$ and $t \in [0, T]$. Then the functional linear model that represents the relationship between $X(s)$ and $Y(t)$ is

$$Y(t) = \alpha(t) + \int_0^T X(s)\beta(s,t)ds + \varepsilon(t), \qquad (1)$$

where $\alpha(t)$ is the intercept function, $\beta(s,t)$ is the coefficient function, and $\varepsilon(t)$ is the error function. Matsui et al. (2009) proposed estimating the model by the basis expansion techniques and the maximum penalized likelihood method and then derived a model selection criterion for evaluating the fitted model. On the other hand, Yao et al. (2005) approached this problem by the functional principal component analysis and showed that their approach is effective for the situation where the longitudinal data are densely or sparsely observed. Furthermore, the comprehensive functional regression model including (1) is developed by Ivanescu et al. (2015); Scheipl et al. (2015, 2016), and they also estimated the model as the mixed effect model framework.

If the arguments s and t in model (1) represent times, and the response depends on future information on the predictor, then it leads to mathematically intractable results. In other words, the integration interval in (1) should be constrained by $s < t$. Malfait and Ramsay (2003) approached this problem by taking the dependence on time between $X(s)$ and $Y(t)$ into consideration and proposed a historical functional linear model (HFLM) as a special case of (1), and they also investigated how to estimate the parameters of the HFLM. Furthermore, Harezlak et al. (2007) estimated the parameters in the HFLM with the penalized least-squares method with the L_2- or the L_1-type penalties. Şentürk and Müller (2008) and Şentürk and Müller (2010) also discussed similar situations for the frameworks of varying-coefficient models. These studies considered functional linear models with a single predictor, while Brockhaus et al. (2017) considered the model with multiple predictors and estimated it by the boosting method.

In this paper, we consider a variable selection problem for multiple functional predictors in the HFLM using sparse regularization. Sparse regularization is one of the most useful tools for variable selection problems and has come to be used in various situations. It can simultaneously estimate parameters and select variables by imposing L_1-type penalties (Bühlmann and van de Geer, 2011; Hastie et al., 2015). For functional linear models with a scalar response, functional predictors are selected using the sparse regularization in Matsui and Konishi (2011), Zhao et al. (2012), Mingotti et al. (2013), and Matsui (2014). We propose a strategy for the problem of variable selection in the historical functional linear model. The functional predictors and the functional response are represented by basis expansions. Since it is difficult to analytically evaluate functions in the model, we apply the finite element method (FEM) introduced by Malfait and Ramsay (2003). Then, the parameters to be included in the model are estimated by the maximum penalized likelihood method with the sparse regularization such as SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), and MCP (Zhang, 2010). Furthermore, we apply the idea of the group lasso (Yuan and Lin, 2006) to these penalties and then shrink each set of parameters that is relevant to the functional variable. In order to decide the degrees of regularization, we apply a model selection criterion (Konishi and Kitagawa, 2008) obtained for evaluating the functional linear model. Monte

Carlo simulations are conducted to assess the effectiveness of the proposed modeling strategy. Then, we apply the proposed method to the analysis of typhoon data to select the functional variables which can explain the trajectories of typhoons.

This paper is organized as follows. Section 2 introduces a HFLM that models the relationship between multiple predictors and a response, both of which are functions of time. Section 3 provides a method for estimating the model parameters using numerical approximation and penalized likelihood method, and then shows a model selection criterion for evaluating the model. Numerical examples are investigated in Section 4 and an example using real data is presented in Section 5. Finally, we summarize the main points in Section 6.

2. Historical functional linear model

Suppose we have n sets of p functional predictors and a functional response $\{(x_{ij}(s), y_i(t));\ s,t \in [0,T],\ i = 1,\ldots,n,\ j = 1,\ldots,p\}$ where s and t represent times. Both functional data $x_{ij}(s)$ and $y_i(t)$ are supposed to be obtained by basis expansions from observed longitudinal data. In order to model the relationship between the predictors and the response, we consider the following historical functional linear model (HFLM, Malfait and Ramsay, 2003; Ramsay and Silverman, 2005):

$$y_i(t) = \alpha(t) + \sum_{j=1}^{p} \int_{s_j(t)}^{t} x_{ij}(s)\beta_j(s,t)ds + e_i(t), \tag{2}$$

where $\alpha(t)$ is an intercept function, $\beta_j(s,t)$ is a bivariate coefficient function which imposes varying weights on $x_{im}(s)$ at $s \in [s_j(t), t]$ rather than $s \in [0,T]$, $s_j(t) = \max\{0, t-\delta_j\}$ with lag parameter $\delta_j > 0$, which determines the lag time up until which variables are included in the model, and $e_i(t)$ are error functions. The HFLM relates to other functional regression models as follows. If intervals of the integration with respect to s are shrunk to $s_j(t) = t$, that is, the arguments of the predictors and the response are the same, the HFLM (2) corresponds to the functional concurrent model:

$$y_i(t) = \alpha(t) + \sum_{j=1}^{p} x_{ij}(t)\beta_j(t) + e_i(t).$$

On the other hand, if $[s_j(t), t]$ is discretized to $t_l,\ l = 1,\ldots,R_j$ so that $t_l = t$ and $t_{l-(R_j-1)} = s_j(t)$ for fixed t, then it corresponds to the generalized varying-coefficient model by Şentürk and Müller (2008) with multiple predictors:

$$y_i(t_l) = \alpha(t_l) + \sum_{j=1}^{p}\sum_{r=1}^{R_j} x_{ij}(t_{l-(r-1)})\beta_{jr}(t_l) + e_i(t_l).$$

Focusing the HFLM (2), from the normal equation, the intercept function is given by

$$\alpha(t) = \bar{y}(t) - \sum_{j=1}^{p} \int_{s_j(t)}^{t} \bar{x}_j(s)\beta_j(s,t)ds + \bar{e}(t),$$

with $\bar{x}_j(s) = \sum_i x_{ij}(s)/n$, $\bar{y}(t) = \sum_i y_i(t)/n$ and $\bar{e}(t) = \sum_i e_i(t)/n$. Therefore, the HFLM (2) becomes

$$y_i^c(t) = \sum_{j=1}^{p} \int_{s_0(t)}^{t} x_{ij}^c(s)\beta_j(s,t)ds + e_i^c(t), \tag{3}$$

where $x_{ij}^c(s) = x_{ij}(s) - \bar{x}_j(s)$, $y_i^c(t) = y_i(t) - \bar{y}(t)$ and $e_i^c(t) = e_i(t) - \bar{e}(t)$. For the rest of this paper we consider the model (3) and we drop the suffix c for simplicity.

Suppose that the coefficient functions $\beta_j(s,t)$ are approximated by basis expansions:

$$\tilde{\beta}_j(s,t) = \sum_{k=1}^{K_j} b_{jk}\phi_{jk}(s,t), \tag{4}$$

where b_{jk} are unknown coefficient parameters and $\beta_j(s,t)$ are two-dimensional basis functions. Radial basis functions, or thin-plate splines, are widely used for two-dimensional basis functions, but here we use another basis whose details are given in the next section. Denoting the residual between $\beta(s,t)$ and its approximation (4) as $\tilde{\varepsilon}_{(j)}(s,t) = \beta_j(s,t) - \tilde{\beta}_j(s,t)$, the HFLM (3) is expressed by

$$\begin{aligned} y_i(t) &= \sum_{j=1}^{p} \int_{s_j(t)}^{t} x_{ij}(s) \left\{ \sum_{k=1}^{K_j} b_{jk}\phi_{jk}(s,t) + \tilde{\varepsilon}_{(j)}(s,t) \right\} ds + e_i(t) \\ &= \sum_{j=1}^{p}\sum_{k=1}^{K_j} b_{jk} \int_{s_j(t)}^{t} x_{ij}(s)\phi_{jk}(s,t)ds + \varepsilon_i(t) \\ &= \sum_{j=1}^{p}\sum_{k=1}^{K_j} b_{jk}\psi_{ijk}(t) + \varepsilon_i(t), \end{aligned}$$

where

$$\psi_{ijk}(t) = \int_{s_j(t)}^{t} x_{ij}(s)\phi_{jk}(s,t)ds$$

and we write the residual $\sum_j \int x_{ij}(s)\tilde{\varepsilon}_{(j)}(t)ds + e_i(t)$ as $\varepsilon_i(t)$ in the last line. Using vector and matrix notations $\boldsymbol{y}(t) = (y_1(t), \ldots, y_n(t))'$, $\boldsymbol{\varepsilon}(t) = (\varepsilon_1(t), \ldots, \varepsilon_n(t))'$, $\Psi_j(t) = (\psi_{ijk}(t))_{ik}$, $\Psi(t) = (\Psi_1(t), \ldots, \Psi_p(t))$, $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_p')'$, and $\boldsymbol{b}_j = (b_{j1}, \ldots, b_{jK_j})'$, the model (3) takes the form

$$\begin{aligned} \boldsymbol{y}(t) &= \sum_{j=1}^{p} \Psi_j(t)\boldsymbol{b}_j + \boldsymbol{\varepsilon}(t) \\ &= \Psi(t)\boldsymbol{b} + \boldsymbol{\varepsilon}(t), \end{aligned} \tag{5}$$

which is similar to the standard linear model with a design matrix $\Psi(t)$, a response vector $\boldsymbol{y}(t)$, and coefficients $\boldsymbol{b}$, except that some of the vectors are functions of t.

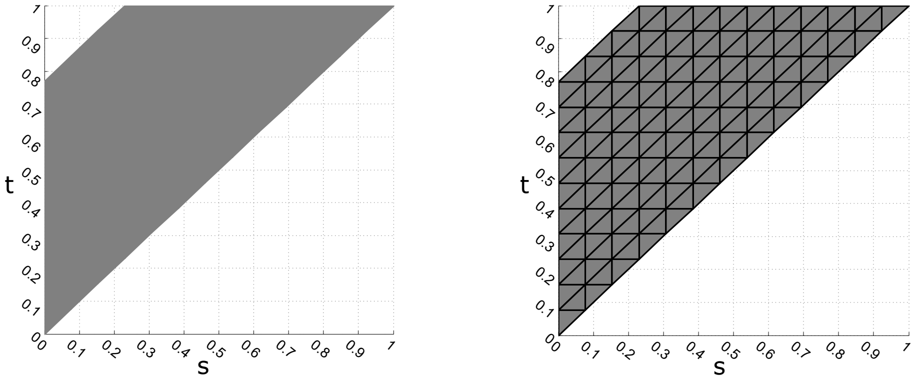

Figure 1: Illustration of the region of the integral of the HFLM (left) and its triangulation (right) for the time lag parameter $\delta_j \in (0, T)$.

3. Estimation and evaluation

We consider the estimation of the functional linear model described in the previous section. As a natural approach for this problem, Ramsay and Silverman (2005), Chapter 16 considered minimizing the integrated sum of squared errors $\int_0^T \sum_{i=1}^n \varepsilon_i^2(t)dt$. However, it is difficult to obtain estimates of $\boldsymbol{b}$ in (5) directly, since it is difficult to calculate ψ_{ijk} analytically due to the complexity of the integral. To solve this problem, we approximate the integral numerically by applying the FEM. In addition, we apply sparse regularization to select the functional predictors and obtain stable estimates.

3.1. Finite element method

In order to calculate ψ_{ijk} numerically, here we apply the FEM to estimate the coefficient vector $\boldsymbol{b}$. Malfait and Ramsay (2003) used the FEM for the HFLM and described the relevant details.

Consider a two-dimensional coordinate system (s, t) which includes the domain of integration in (2), as displayed in Figure 1, left. First we divide the intervals $[0, T]$ for s and t directions into N equally spaced intervals with length μ, and then construct triangular elements by further dividing each square grid into two triangles, as shown in Figure 1, right. The value of δ_j can be approximated by $M_j\mu$ $(0 \leq M_j \leq N)$ for each j. When $M_j = N$, the domain becomes a triangle and $0 < M_j < N$ corresponds to a trapezoid (as depicted in Figure 1 left), and when $M_j = 0$, the domain is $s = t$, which corresponds to the functional concurrent model. As a result, there are $T_j = M_j(2N - M_j)$ triangular elements and $V_j = (M_j + 1)(N + 1 - M_j/2)$ nodes on the domain of $\beta_j(s, t)$. Each node is assigned by one basis function which has the shape of a hexagonal pyramid and has a value of 1 at the node and 0 at the adjacent nodes, and therefore $K_j = V_j$. These bases correspond to $\phi_{jk}(t)$ $(j = 1, \ldots, p,\ k = 1, \ldots, K_j)$ in equation (4). We used Matlab functions to calculate the triangulation, available from the website of Ramsay and Silverman (2002).

We consequently discretize the time t into Q finite time points . Malfait and Ramsay (2003) showed that $Q = 4N$ gives sufficient accuracy for the approximation. Using this discretization, the vector forms of $y_i(t)$, $\psi_{ijk}(t)$, and $\varepsilon_i(t)$ are $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iQ})'$,

$\boldsymbol{\psi}_{ijk} = (\psi_{i1jk}, \ldots, \psi_{iQjk})'$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iQ})'$. Then, using the notation

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix}, \Psi = \begin{pmatrix} \boldsymbol{\psi}_{111} & \cdots & \boldsymbol{\psi}_{11K_1} & \cdots & \boldsymbol{\psi}_{1p1} & \cdots & \boldsymbol{\psi}_{1pK_p} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{\psi}_{n11} & \cdots & \boldsymbol{\psi}_{n1K_1} & \cdots & \boldsymbol{\psi}_{np1} & \cdots & \boldsymbol{\psi}_{npK_p} \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

we can represent the discretized version of HFLM as

$$\boldsymbol{y} = \Psi \boldsymbol{b} + \boldsymbol{\varepsilon}. \tag{6}$$

An advantage of applying the FEM is that we can approximate the HFLM (3) with (6), the same form as the traditional linear model. Then we can use several estimation procedures which are applied to the linear model.

3.2. Penalized likelihood method via sparse regularization

We assume that the error vectors $\boldsymbol{\varepsilon}_i$ are identically and independently normally distributed with mean vector $\mathbf{0}$ and variance-covariance matrix Σ_0, and that Σ_0 has an autocorrelation structure, since $\varepsilon_{i1}, \ldots, \varepsilon_{iQ}$ are discretized realizations of the error function over time. That is, we assume that Σ_0 has the form

$$\Sigma_0 = \frac{\sigma^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \cdots & \rho^{Q-1} \\ \rho & 1 & \cdots & \rho^{Q-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{Q-1} & \rho^{Q-2} & \cdots & 1 \end{pmatrix}, \tag{7}$$

where σ^2 and $\rho \in [-1, 1]$ are variance and correlation parameters, respectively. Here we referred to Fahrmeir et al. (2013) for the covariance structure. The variance-covariance matrix of $\boldsymbol{\varepsilon}$ is given by $\Sigma = I_n \otimes \Sigma_0$, and hence we have probability density function

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = \frac{1}{(2\pi)^{nQ/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \Psi \boldsymbol{b})' \Sigma^{-1} (\boldsymbol{y} - \Psi \boldsymbol{b}) \right\}, \tag{8}$$

where $\boldsymbol{\theta} = (\boldsymbol{b}', \sigma^2, \rho)'$ is a vector of the parameters. The log-likelihood function is given by

$$\ell(\boldsymbol{\theta}) = -\frac{nQ}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (\boldsymbol{y} - \Psi \boldsymbol{b})' \Sigma^{-1} (\boldsymbol{y} - \Psi \boldsymbol{b}).$$

We estimate parameters $\boldsymbol{b}$, σ^2, and ρ by maximizing the following penalized likelihood function

$$\ell_\lambda(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - nQ \sum_{j=1}^{p} P_\lambda(\|\boldsymbol{b}_j\|_{\Omega_j}), \tag{9}$$

where $P_\lambda(\cdot)$ is a penalty function which is controlled by a regularization parameter $\lambda > 0$ and $\|\boldsymbol{b}_j\|_{\Omega_j} = \sqrt{\boldsymbol{b}_j' \Omega_j \boldsymbol{b}_j}$ with positive semi-definite matrix Ω_j. We used $P_\lambda(\|\boldsymbol{b}_j\|_{\Omega_j})$ rather than $P_\lambda(|b_{jk}|)$ as the penalty for the penalized likelihood method by applying the idea of the group lasso, as we can shrink all of the parameters relevant to the j-th predictor toward zero simultaneously. In particular, here, we consider SCAD, elastic net and the MCP for the panlty function $P_\lambda(\cdot)$. Furthermore, Harezlak et al. (2007) penalized fluctuations of the coefficient function in three directions: parallel to the s-axis (horizontal), t-axis (vertical), and $s = t$ (parallel), which respectively corresponds to penalizing coefficients for neighboring basis functions for fixed t, s, and $s - t$ (Figure 2). We used the first differences of the coefficients of the neighboring bases by constructing penalty matrices for the horizontal, vertical, and parallel directions for the j-th variable, respectively denoted by $D_j^{(H)}$, $D_j^{(V)}$, and $D_j^{(P)}$. For example, the elements of $D_j^{(H)}$ are given by

$$\begin{cases} \left(D_j^{(H)}\right)_{lv} = 1 \\ \left(D_j^{(H)}\right)_{lv'} = -1 \\ \left(D_j^{(H)}\right)_{lv''} = 0 \end{cases} \left(\text{if} \quad \begin{array}{c} s_{(v)} - s_{(v')} = 1, \\ t_{(v)} - t_{(v')} = 0, \\ v'' \neq v, v' \end{array} \right), \tag{10}$$

where $s_{(v)}$ and $t_{(v)}$ are the s and t coordinate values of the v-th node, respectively, $v = 1, \dots, V_j$, and $l = 1, \dots, L$, with L being the number of combinations where the condition in (10) is satisfied. The matrices $D_j^{(V)}$ and $D_j^{(P)}$ are given in similar ways. Then the matrices Ω_j are given by

$$\Omega_j = \gamma_j^{(H)} (D_j^{(H)})' D_j^{(H)} + \gamma_j^{(V)} (D_j^{(V)})' D_j^{(V)} + \gamma_j^{(P)} (D_j^{(P)})' D_j^{(P)}$$

with tuning parameters $\gamma_j^{(H)}$, $\gamma_j^{(V)}$, and $\gamma_j^{(P)}$ that control the degrees of regularization for each direction. Although we can select all of these parameters by some model selection criteria, this can be computationally expensive. Instead, we decide these values with the following rule, using the idea of Fan and Li (2004). First, we obtain the maximum likelihood estimator of $\boldsymbol{b}$, denoted by $\hat{\boldsymbol{b}}^{(ML)}$, by maximizing (9) with $\lambda = 0$, and then $\gamma_j^{(H)}$ is given as standard deviations of $D_j^{(H)} \hat{\boldsymbol{b}}_j^{(ML)}$. The other parameters $\gamma_j^{(V)}$ and $\gamma_j^{(P)}$ are obtained in the same way, and then λ is selected using the model selection criteria introduced in the next subsection. When it is difficult to derive $\hat{\boldsymbol{b}}^{(ML)}$ (e.g. the dimension of $\boldsymbol{b}$ exceeds the sample size n), we instead use the maximum likelihood estimator with generalized inverse or penalized maximum likelihood estimator with small regularization parameter.

The parameters to be included in the model are estimated by the local quadratic approximation, which iteratively updates the parameter estimates and has been applied for the nonconcave penalties such as SCAD and MCP. For the concave penalty including the elastic net, several other optimization algorithms are proposed such as coordinate descent algorithm and the augmented Lagrangian method, but the quadratic approximation can be also applied. Denoting the initial value of $\boldsymbol{b}$ in the updated estimate as $\boldsymbol{b}^{(0)}$, we approximated the penalty function $P_\lambda(\|\boldsymbol{b}_j\|_{\Omega_j})$ by

$$P_\lambda(\|\boldsymbol{b}_j\|_{\Omega_j}) \approx P_\lambda(\|\boldsymbol{b}_j^{(0)}\|_{\Omega_j}) + \frac{1}{2} \frac{P'_\lambda(\|\boldsymbol{b}_j^{(0)}\|_{\Omega_j})}{\|\boldsymbol{b}_j^{(0)}\|_{\Omega_j}} \left\{ \boldsymbol{b}_j' \boldsymbol{b}_j - \left(\boldsymbol{b}_j^{(0)}\right)' \boldsymbol{b}_j^{(0)} \right\}.$$

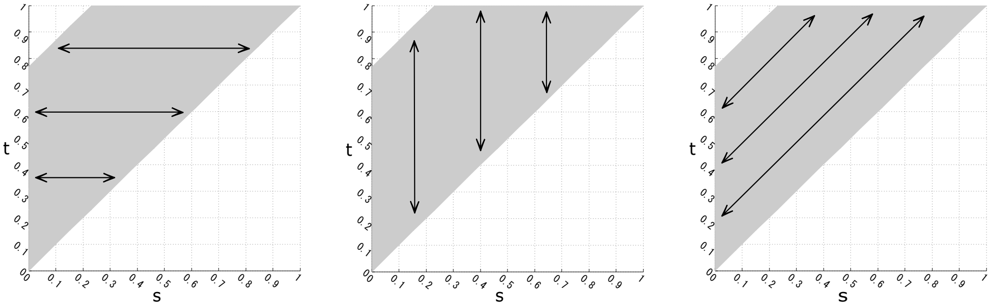

Figure 2: Directions for penalizations about $D_j^{(H)}$ (left), $D_j^{(V)}$ (center) and $D_j^{(P)}$ (right).

Using this approximation and assuming that σ^2 and ρ are fixed, the penalized log-likelihood function (9) is approximated by the Taylor expansion:

$$\ell_\lambda(\boldsymbol{b}) \approx \ell(\boldsymbol{b}^{(0)}) + \frac{\partial \ell(\boldsymbol{b}^{(0)})}{\partial \boldsymbol{b}'}(\boldsymbol{b}-\boldsymbol{b}^{(0)}) + \frac{1}{2}(\boldsymbol{b}-\boldsymbol{b}^{(0)})' \frac{\partial \ell(\boldsymbol{b}^{(0)})}{\partial \boldsymbol{b} \partial \boldsymbol{b}'}(\boldsymbol{b}-\boldsymbol{b}^{(0)}) + \frac{nQ}{2}\boldsymbol{b}'\Omega_\lambda(\boldsymbol{b})\boldsymbol{b},$$

where $\Omega_\lambda(\boldsymbol{b}) = \mathrm{blockdiag}\{P'_\lambda(\|\boldsymbol{b}_1\|_{\Omega_1})/\|\boldsymbol{b}_1\|_{\Omega_1}, \ldots, P'_\lambda(\|\boldsymbol{b}_p\|_{\Omega_p})/\|\boldsymbol{b}_p\|_{\Omega_p}\}$. Then if the k-th updated values of $\boldsymbol{b}$ and Σ, denoted by $\boldsymbol{b}^{(k)}$ and $\Sigma^{(k)}$, respectively, are obtained, the $(m+1)$-th updated value of $\boldsymbol{b}$ is given by

$$\boldsymbol{b}^{(m+1)} = \left\{\Psi' \left(\Sigma^{(k)}\right)^{-1} \Psi + nQ\Omega_\lambda(\boldsymbol{b}^{(k)})\right\}^{-1} \Psi' \left(\Sigma^{(k)}\right)^{-1} \boldsymbol{y}, \tag{11}$$

and subsequently the correlation and variance parameters in (7) are updated by

$$\rho^{(m+1)} = \frac{s_{q1}^{(m+1)}}{s_q^{(m+1)}}, \quad \left(\sigma^2\right)^{(m+1)} = \frac{1}{nQ}\left(\boldsymbol{y} - \Psi \boldsymbol{b}^{(m+1)}\right)' \left(\boldsymbol{y} - \Psi \hat{\boldsymbol{b}}^{(m+1)}\right), \tag{12}$$

respectively, where

$$s_{q1}^{(m+1)} = \frac{1}{nQ}\sum_{i=1}^{n}\sum_{q=2}^{Q}\left(y_{iq} - \sum_{j=1}^{p}\sum_{k=1}^{K_j}\psi_{iqjk}b_{jk}^{(m+1)}\right)\left(y_{i(q-1)} - \sum_{j=1}^{p}\sum_{k=1}^{K_j}\psi_{i(q-1)jk}b_{jk}^{(m+1)}\right),$$

$$s_{q}^{(m+1)} = \frac{1}{nQ}\sum_{i=1}^{n}\sum_{q=1}^{Q}\left(y_{iq} - \sum_{j=1}^{p}\sum_{k=1}^{K_j}\psi_{iqjk}b_{jk}^{(m+1)}\right)^2.$$

The updated variance-covariance matrix $\Sigma^{(m+1)}$ is obtained by substituting $\rho^{(m+1)}$ and $(\sigma^2)^{(m+1)}$ into (7). This algorithm is summarized as follows:

1. Choose initial values of parameters $\boldsymbol{b}^{(0)}$, $\rho^{(0)}$, and $(\sigma^2)^{(0)}$.

2. Update the coefficients $\boldsymbol{b}$ by (11).

3. Update the parameters ρ and σ^2 in the variance-covariance matrix Σ by (12).

4. Repeat steps 2 and 3 until convergence.

Substituting the estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{b}}', \hat{\sigma}^2, \hat{\rho})'$ into (8), we obtain the following model:

$$f(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \frac{1}{(2\pi)^{nQ/2}|\hat{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y} - \Psi\hat{\boldsymbol{b}}\right)' \hat{\Sigma}^{-1}\left(\boldsymbol{y} - \Psi\hat{\boldsymbol{b}}\right)\right\}. \tag{13}$$

3.3. Model selection criterion

As described in the previous section, the model defined in (13) depends on the regularization parameter λ, and therefore we need to select an appropriate value of it. Although cross-validation is widely used for the selection of tuning parameters, it tends to be computationally expensive. Furthermore, Leng et al. (2006) showed that criteria based on the minimum prediction error, such as cross-validation or generalized cross-validation, are not consistent. On the other hand, Wang et al. (2007) and Zhang et al. (2010) showed that the Bayesian information criterion (BIC) with the effective degrees of freedom is consistent for the SCAD regularization. We used a BIC-type model selection criterion for evaluating the historical functional linear model (2) estimated by the maximum penalized likelihood method with the sparse regularization. The BIC is given by

$$\mathrm{BIC} = -2\ell(\hat{\boldsymbol{\theta}}) + \mathit{df}\log(nQ),$$

where

$$\mathit{df} = \mathrm{tr}\left\{\Psi\left(\Psi'\hat{\Sigma}^{-1}\Psi + nQ\Omega(\hat{\boldsymbol{b}})\right)^{-1}\Psi'\hat{\Sigma}^{-1}\right\}$$

is an effective degrees of freedom derived using the idea of Hastie and Tibshirani (1990). Note that the part of the formula of the effective degrees of freedom is the same as (11) in the LQA algorithm, and therefore the BIC can be easily calculated. We select the regularization parameter λ so as to minimize the BIC and then consider this to be an optimal model.

4. Simulation study

We conducted Monte Carlo simulations to show the effectiveness of the proposed method. We simulated predictors $X_j(s)$ $(j = 1, \ldots, 5)$ and a response $Y(t)$ with $s, t \in [0, 1]$ in the HFLM. First, we generated functional predictors and coefficients by

$$x_{ij}(s) = \sum_{k=1}^{K}(u_{jk} + w_{ijk})\xi_{jk}(s), \quad \beta_j(s,t) = \sum_{k=1}^{K^2} v_{jk}\xi_{jk}(s)\xi_{jk}(t),$$

where $K = 7$ and $\xi_{jk}(s)$ are the basis functions. Here, we applied the Gaussian radial basis functions (Bishop, 1995) for $\xi_{jk}(s)$. An advantage with respect to the use of Gaussian bases is that it is straightforward to simulate the true response function (Matsui and Konishi, 2011). Furthermore, coefficients u_{jk} and w_{ijk} in $x_{ij}(s)$ are given by

$$u_{1k} = k, \quad u_{2k} = \sin(3\pi k), \quad u_{3k} = \cos(3\pi k), \quad u_{4k} = \sin(5\pi k), \quad u_{5k} = \cos(5\pi k),$$
$$\boldsymbol{w}_{ij} \sim N_K(\boldsymbol{0}, \Sigma_w), \quad \boldsymbol{w}_{ij} = (w_{ij1}, \ldots, w_{ijK})', \quad \Sigma_w = (\sigma_w \rho_w^{|k-l|})_{kl},$$

respectively, where $\sigma_w = 0.3$ and $\rho_w = 0.5$, and v_{jk} in $\beta_j(s,t)$ are set as

$$v_{1k} = 0.5^{(a-4)^2+(b-4)^2}/10, \quad v_{2k} = 0.1a, \quad v_{3k} = -\sin((a-b)^2), \quad v_{4k} = 0, \quad v_{5k} = 0,$$

where $a = 1, \dots, K$ and $b = 1, \dots, K$ are given so that $k = (a-1)K + b$. This setting means that the variables X_4 and X_5 are unnecessary for the model. Furthermore, the error function in the HFLM was generated by the basis expansion with random coefficients:

$$\varepsilon_i(t) = \sum_{k=1}^{K} \varepsilon_k \xi_k(t), \quad \varepsilon_k \sim N(0, \sigma_e^2 R_i^{e2}),$$

where $\xi_k(t)$ are the same basis functions as $\xi_{jk}(t)$, $\sigma_e = 0.05, 0.1$ and $R_i^e = \mathrm{sd}(y_i(t))$ with standard deviation $\mathrm{sd}(\cdot)$. The response is given by

$$y_i(t) = \sum_{j=1}^{5} \int_{s_j(t)}^{t} x_{ij}(s)\beta_j(s,t)ds + \varepsilon_i(t),$$

where we set the true lag parameter δ_j included in $s_j(t)$ to be $\delta_j = 0.5$ for all j. Since the observed data contain noise, we added noise to the above predictors and responses as follows:

$$x_{ijl} = x_{ij}(s_l) + \varepsilon_{ijl}^{(x)}, \quad y_{il} = y_i(t_l) + \varepsilon_{il}^{(y)},$$

where $l = 1, \dots, 51$, and $\varepsilon_{ijl}^{(x)}$ and $\varepsilon_{il}^{(y)}$ follow $N(0, \sigma_x^2 R_i^{x2})$ and $N(0, \sigma_y^2 R_i^{y2})$, respectively, with $\sigma_x = \sigma_y = 0.3$, $R_i^x = \mathrm{sd}(x_{ij}(s))$, and $R_i^y = \mathrm{sd}(y_i(t))$.

We treated x_{ijl} and y_{il} as observations, then smoothed them into functional data with B-splines with 8 basis functions using the Matlab functions provided in Ramsay and Silverman (2002). Next we constructed a design matrix and a response vector in (6) using the FEM. There are several tuning parameters involved in the FEM, described in Section 3.1. We set the parameters to $N = 13$ and $\mu = 4$, and M_j is defined such that $\delta_j = 0.25, 0.50, 0.75$ for all j (while the true value is $\delta_j = 0.50$). Then the parameters included in the model were estimated by the maximum likelihood method and the penalized likelihood method with SCAD, elastic net and the MCP. Regularization parameter λ is then selected by the BIC. We conducted this strategy for 100 repetitions and for all combinations of $n = 50, 100$, $\sigma_e = 0.05, 0.1$, and $\delta_j = 0.25, 0.50, 0.75$, and then investigated averaged values of the 100 mean squared errors (MSE) for the response $y_i(t)$ defined by

$$\mathrm{MSE}_y = \frac{1}{51n} \sum_{i=1}^{n} \sum_{l=1}^{51} \left\{ g_i(t_l) - \hat{y}_i(t_l) \right\}^2,$$

where $g_i(t) = y_i(t) - \varepsilon_i(t)$ and $\hat{y}_i(t)$ is an estimated response function. We also investigated the rates of selected variables.

Tables 1 and 2 show the results of the simulation study. Values in the parentheses in these tables are standard deviations of the 100 MSEs and $X_1, \dots, X_5$ indicate the rates of the selected variables. These results demonstrate that, if the lag parameter δ is smaller than the true value (i.e. $\delta = 0.25$) the MLE minimizes the MSEs, but if δ is equal or larger than the true value (i.e. $\delta = 0.50, 0.75$) the MCP gives smaller or competitive MSEs. SCAD gives similar results with MCP, while the elastic net gives larger MSEs for all cases. SCAD and MCP tend to select relatively correct variables, especially for larger n and true δ, but when $\delta = 0.75$ they tend to select unnecessary variables. On the other hand, the elastic net tends to shrink unnecessary variables but it also shrink X_1, the necessary variable. The elastic net selected smaller λs than SCAD and MCP, whereas it gives smaller models than the others. This is probably because the elastic net uniformly imposes penalty to the coefficients, while SCAD and MCP imposes smaller penalties to the larger coefficients due to their nonconcavity. In addition, a possible reason why the elastic net gives worse result is that the BIC with the elastic net does not have consistency for variable selection, which is proved under the oracle property (Wang et al., 2007).

5. Application to real data

We applied the proposed method to the analysis of typhoon data. We investigated which sets of variables in the data influence the path of the typhoons, using the historical functional linear model.

The data are available on the website "Digital Typhoon.[1]" We analyzed 88 typhoons which passed around Japan (30 and 50° N and 130 and 150° E) between 2001 and 2012. The data contain the positions (latitude Y_1 and longitude Y_2), the central atmospheric pressure (X_1), the velocity of the wind around the center (X_2), the radii of major and minor storm axes (areas where the velocities of the winds are higher than 25 m/s; X_3 and X_4, respectively) and those of the major and minor gale axes (areas where the velocities of the winds are higher than 15 m/s; X_5 and X_6, respectively) of the typhoons. These variables were observed every six hours from the formation to the dissipation of the typhoons. Since the survival periods of them differ, we grouped all of the formation and dissipation times of the typhoons as 0 and 1, respectively, by scaling the time points. Figure 3 shows 10 examples of the typhoon data. Since the positions of the time points differ for each subject, it is difficult to apply the traditional linear model directly. On the other hand, we can easily analyze them with functional data analysis. The objective of this analysis is to examine which combination of variables, such as pressure and wind velocity, relate to the location of the typhoon. In order to do this we treated the positions as responses and the other variables as predictors in the functional linear model.

We applied the smoothing method with basis expansions to the observed data and then obtained the functions. Examples of the functions are depicted in Figure 4. After centering the data, we constructed the HFLM (2) by treating the longitude or the latitude of the typhoons as a functional response and the remaining data as functional predictors. Since the model (2) contains only one response, whereas there are two response variables for the position (latitude and longitude), we constructed two HFLMs, one for each of the spatial dimensions. The unknown parameters in the model were estimated by the maximum penalized likelihood method with MCP, and regularization parameter λ in the penalized log-likelihood function was selected with BIC. Here the value of δ_j was fixed at 0.50 for all j. We then investigated the selected variables.

[1] http://agora.ex.nii.ac.jp/digital-typhoon/index.html.en

Figure 6 shows the estimated coefficient functions when the responses are Y_1 and Y_2, respectively. From these figures, we can see that some of the coefficient surfaces were estimated to be zero functions, which leads to the exclusion of the corresponding variables from the model. When the response is the latitude, the coefficient function for the velocity of the wind (X_2) is estimated to be a zero function; that is, only this variable does not affect the latitudinal direction. When the response is the longitude, on the other hand, the predictors about the major and minor storm axes (X_3, X_4) are selected. Therefore, only the storm axes relate to the trajectories of the longitude of the typhoon among the variables.

6. Concluding remarks

We have proposed a method for variable selection in functional linear models where the predictors and the response are functions. When the data are functions of time we need to take into account the dependence in time between the predictors and the response, and therefore we applied the historical functional linear model. Unknown parameters included in the model are estimated by the maximum penalized likelihood method with sparsity inducing penalties, and the regularization parameter was selected by a BIC-type model selection criterion. Simulation results revealed that the proposed method tends to correctly select the predictors under the condition that the lag parameter is equal or smaller than true value. We also applied our method to typhoon data, to determine the combination of variables that best explained the path of the typhoon.

The proposed method can be applied to the dataset with even more variables than that we have applied in this paper. The investigation of the behavior for the case when the number of variables increases remains as a future work. More recently, several numerical algorithms are proposed for nonconcave penalties, such as local linear approximation (Zou and Li, 2008) and alternating direction method of multipliers (Wang et al., 2019). We need more investigations about the application of these algorithms. In addition, we assumed that several tuning parameters, not including regularization parameter λ, were fixed. In particular, we think that it is crucial to select the value of the lag parameter δ_j objectively, since it determines until what time variables are included in the model. However, we cannot directly apply the model selection criteria since the “sample size” in model (6) changes as δ_j changes. Therefore, the selection of δ_j is also the future work. Furthermore, we can consider extending the HFLM to multiple response variables. For example, the longitudes and latitudes in the typhoon data in Section 5 can be modeled simultaneously by taking the correlation among responses into consideration.

Acknowledgement

The author would like to thank the reviewers and the editor for constructive and helpful comments that improved the quality of the paper considerably. This work was supported by Grant-in-Aid for Scientific Research (C) No. 19K11858 of JSPS.

References

Bishop, C. M. (1995), *Neural networks for pattern recognition*, New York: Oxford Univ. Press.

Brockhaus, S., Melcher, M., Leisch, F., and Greven, S. (2017), Boosting flexible functional regression models with a high number of functional historical effects, *Stat. Comput.*, 27, 913–926.

Bühlmann, P. and van de Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Heidelberg: Springer.

Şentürk, D. and Müller, H. (2010), Functional varying coefficient models for longitudinal data, *J. Am. Stat. Assoc.*, 105, 1256–1264.

Şentürk, D. and Müller, H.-G. (2008), Generalized varying coefficient models for longitudinal data, *Biometrika*, 95, 653–666.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013), *Regression*, Berlin Heidelberg: Springer.

Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Stat. Assoc.*, 96, 1348–1360.

— (2004), New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis, *J. Am. Stat. Assoc.*, 99, 710–723.

Fan, Y., James, G. M., and Radchenko, P. (2015), Functional additive regression, *Ann. Statist.*, 43, 2296–2325.

Harezlak, J., Coull, B., Laird, N., Magari, S., and Christiani, D. (2007), Penalized solutions to functional regression problems, *Comput. Statist. Data An.*, 51, 4911–4925.

Hastie, T. and Tibshirani, R. (1990), *Generalized additive models*, London: Chapman & Hall/CRC.

— (1993), Varying-coefficient models, *J. Roy. Statist. Soc. B*, 55, 757–796.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalization*, Boca Raton: Chapman & Hall/CRC.

Hoover, D., Rice, J., Wu, C., and Yang, L. (1998), Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, *Biometrika*, 85, 809–822.

Ivanescu, A. E., Staicu, A. M., Scheipl, F., and Greven, S. (2015), Penalized function-on-function regression, *Comput. Statist.*, 30, 539–568.

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

James, G. and Silverman, B. (2005), Functional adaptive model estimation, *J. Am. Stat. Assoc.*, 100, 565–576.

Kokoszka, P. and Reimherr, M. (2017), *Introduction to functional data analysis*, Boca Raton: CRC Press.

Konishi, S. and Kitagawa, G. (2008), *Information criteria and statistical modeling*, New York: Springer.

Leng, C., Lin, Y., and Wahba, G. (2006), A note on the lasso and related procedures in model selection, *Statist. Sinica*, 16, 1273–1284.

Malfait, N. and Ramsay, J. (2003), The historical functional linear model, *Can. J. Statist.*, 31, 115–128.

Matsui, H. (2014), Variable and boundary selection for functional data via multiclass logistic regression modeling, *Comput. Statist. Data An.*, 78, 176–185.

Matsui, H., Kawano, S., and Konishi, S. (2009), Regularized functional regression modeling for functional response and predictors, *J. Math-for-Indust.*, 1, 17–25.

Matsui, H. and Konishi, S. (2011), Variable selection for functional regression models via the L1 regularization, *Comput. Statist. Data An.*, 55, 3304–3310.

Mingotti, N., Lillo, R., and Romo, J. (2013), Lasso variable selection in functional regression, *Statistics and Econometrics Working Papers from Universidad Carlos III.*

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

Müller, H.-G. and Yao, F. (2008), Functional Additive Models, *J. Am. Stat. Assoc.*, 103, 1534–1544.

Ramsay, J. and Dalzell, C. (1991), Some tools for functional data analysis, *J. Roy. Statist. Soc. B*, 53, 539–572.

Ramsay, J. and Silverman, B. (2002), *Applied functional data analysis: methods and case studies*, New York: Springer.

— (2005), *Functional data analysis (2nd ed.)*, New York: Springer.

Scheipl, F., Gertheiss, J., and Greven, S. (2016), Generalized functional additive mixed models, *Electron. J. Statist.*, 10, 1455–1492.

Scheipl, F., Staicu, A.-M., and Greven, S. (2015), Functional Additive Mixed Models, *J. Comput. Graph. Stat.*, 24, 477–501.

Wang, H., Li, R., and Tsai, C. (2007), Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika*, 94, 553–568.

Wang, Y., Yin, W., and Zeng, J. (2019). Global convergence of ADMM in nonconvex nonsmooth optimization, *J. Sci. Comput.*, 78, 29–63.

Yao, F., Müller, H., and Wang, J. (2005), Functional linear regression analysis for longitudinal data, *Ann. Statist.*, 33, 2873–2903.

Yuan, M. and Lin, Y. (2006), Model selection and estimation in regression with grouped variables, *J. Roy. Statist. Soc. B*, 68, 49–67.

Zhang, C. (2010), Nearly unbiased variable selection under minimax concave penalty, *Ann. Statist.*, 38, 894–942.

Zhang, Y., Li, R., and Tsai, C. (2010), Regularization parameter selections via generalized information criterion, *J. Am. Stat. Assoc.*, 105, 312–323.

Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012), Wavelet-based LASSO in functional linear regression, *J. Comput. Graph. Stat.*, 21, 600–617.

Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net, *J. Roy. Statist. Soc. B*, 67, 301–320.

Zou, H., and Li, R. (2008), One-step sparse estimates in nonconcave penalized likelihood models, *Ann. Statist.*, 36(4), 1509–1533.

Received: October 12, 2020
Revised: December 29, 2020
Accept: January 6, 2021

Table 1: Results on 100 repetitions in simulation studies for $n = 50$.

$\sigma_e = 0.05, \delta = 0.25$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	2.97 (0.14)	0	1.00	1.00	1.00	1.00	1.00
SCAD	3.02 (0.18)	1.00×10^{-2}	0.91	1.00	1.00	0.26	0.31
Elastic net	11.12 (0.95)	3.13×10^{-5}	0.04	1.00	1.00	0.00	0.00
MCP	3.16 (0.30)	6.44×10^{-3}	0.60	1.00	1.00	0.43	0.44

$\sigma_e = 0.1, \delta = 0.25$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	3.18 (0.14)	0	1.00	1.00	1.00	1.00	1.00
SCAD	3.19 (0.22)	1.02×10^{-2}	0.92	1.00	1.00	0.43	0.43
Elastic net	11.15 (1.14)	3.63×10^{-5}	0.02	1.00	1.00	0.02	0.02
MCP	3.48 (0.34)	8.22×10^{-3}	0.42	1.00	1.00	0.27	0.28

$\sigma_e = 0.05, \delta = 0.50$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	2.14 (0.12)	0	1.00	1.00	1.00	1.00	1.00
SCAD	2.07 (0.16)	6.31×10^{-3}	0.96	1.00	1.00	0.37	0.45
Elastic net	11.16 (1.05)	3.16×10^{-5}	0.07	1.00	1.00	0.00	0.03
MCP	2.07 (0.17)	3.09×10^{-3}	0.94	1.00	1.00	0.27	0.33

$\sigma_e = 0.1, \delta = 0.50$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	2.51 (0.15)	0	1.00	1.00	1.00	1.00	1.00
SCAD	2.55 (0.20)	4.46×10^{-3}	0.89	1.00	1.00	0.44	0.45
Elastic net	11.10 (0.92)	3.06×10^{-5}	0.03	1.00	1.00	0.00	0.01
MCP	2.43 (0.19)	3.56×10^{-2}	0.43	1.00	1.00	0.29	0.27

$\sigma_e = 0.05, \delta = 0.75$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	2.21 (0.13)	0	1.00	1.00	1.00	1.00	1.00
SCAD	2.19 (0.14)	5.01×10^{-3}	1.00	1.00	1.00	0.88	0.86
Elastic net	11.15 (1.04)	2.76×10^{-5}	0.04	1.00	1.00	0.02	0.05
MCP	2.16 (0.20)	3.29×10^{-3}	0.95	1.00	1.00	0.52	0.50

$\sigma_e = 0.1, \delta = 0.75$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	2.63 (0.18)	0	1.00	1.00	1.00	1.00	1.00
SCAD	2.63 (0.19)	3.30×10^{-3}	0.85	1.00	1.00	0.82	0.81
Elastic net	10.89 (0.97)	3.16×10^{-5}	0.06	1.00	1.00	0.03	0.01
MCP	2.52 (0.22)	3.26×10^{-3}	0.99	1.00	1.00	0.62	0.58

Table 2: Results on 100 repetitions in simulation studies for $n = 100$.

$\sigma_e = 0.05, \delta = 0.25$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	3.03 (0.12)	0	1.00	1.00	1.00	1.00	1.00
SCAD	3.68 (0.20)	1.02×10^{-2}	0.00	1.00	1.00	0.00	0.00
Elastic net	11.28 (0.64)	3.54×10^{-5}	0.00	1.00	1.00	0.00	0.00
MCP	3.06 (0.12)	3.10×10^{-3}	0.99	1.00	1.00	0.17	0.09

$\sigma_e = 0.1, \delta = 0.25$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	3.12 (0.11)	0	1.00	1.00	1.00	1.00	1.00
SCAD	3.24 (0.25)	6.31×10^{-2}	0.74	1.00	1.00	0.09	0.02
Elastic net	11.32 (0.65)	3.24×10^{-5}	0.00	1.00	1.00	0.00	0.00
MCP	3.13 (0.13)	3.19×10^{-3}	0.97	1.00	1.00	0.22	0.11

$\sigma_e = 0.05, \delta = 0.50$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	1.97 (0.09)	0	1.00	1.00	1.00	1.00	1.00
SCAD	1.93 (0.12)	3.16×10^{-3}	0.99	1.00	1.00	0.29	0.31
Elastic net	11.18 (0.69)	3.24×10^{-5}	0.00	1.00	1.00	0.00	0.00
MCP	1.92 (0.10)	1.58×10^{-3}	1.00	1.00	1.00	0.23	0.26

$\sigma_e = 0.1, \delta = 0.50$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	2.18 (0.08)	0	1.00	1.00	1.00	1.00	1.00
SCAD	2.15 (0.09)	3.14×10^{-3}	0.78	1.00	1.00	0.73	0.75
Elastic net	11.22 (0.72)	3.16×10^{-5}	0.00	1.00	1.00	0.00	0.00
MCP	2.22 (0.18)	1.37×10^{-3}	0.96	1.00	1.00	0.23	0.10

$\sigma_e = 0.05, \delta = 0.75$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	1.99 (0.09)	0	1.00	1.00	1.00	1.00	1.00
SCAD	2.80 (1.83)	3.64×10^{-3}	0.94	1.00	0.83	0.33	0.31
Elastic net	11.09 (0.64)	3.32×10^{-5}	0.00	1.00	1.00	0.00	0.00
MCP	1.99 (0.09)	1.12×10^{-3}	1.00	1.00	1.00	0.92	0.93

$\sigma_e = 0.1, \delta = 0.75$

	MSE	λ	X_1	X_2	X_3	X_4	X_5
MLE	2.17 (0.13)	0	1.00	1.00	1.00	1.00	1.00
SCAD	3.18 (2.13)	3.33×10^{-3}	1.00	1.00	0.90	0.60	0.70
Elastic net	11.99 (0.52)	5.44×10^{-5}	0.00	1.00	1.00	0.00	0.00
MCP	2.13 (0.09)	1.58×10^{-3}	1.00	1.00	1.00	1.00	1.00

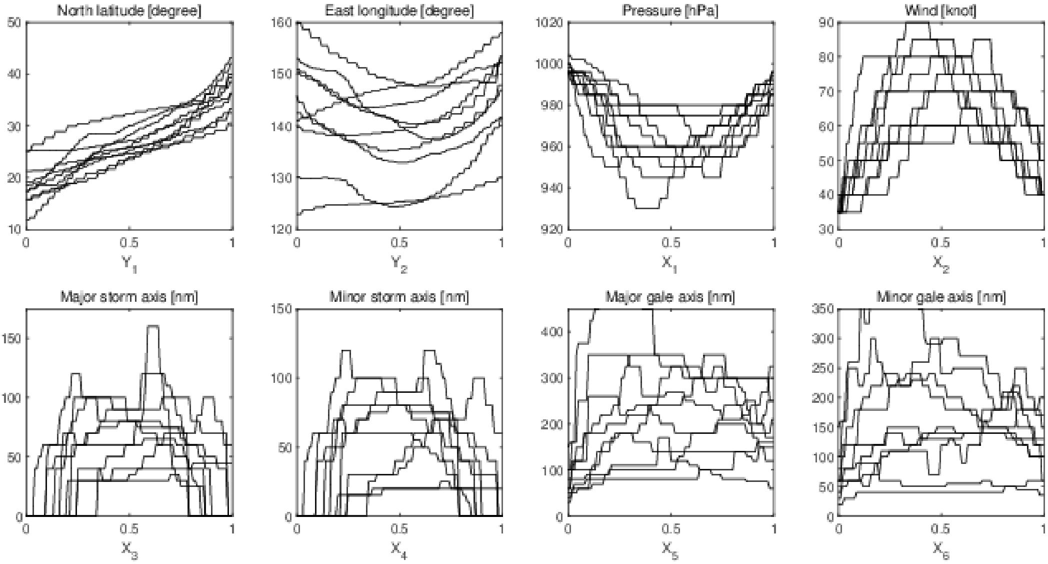

Figure 3: Examples of typhoon data. Two plots in the top left (north latitude and east longitude) are responses and the remaining data are predictors. The square brackets indicates units of measurement, where "nm" represents the nautical mile.

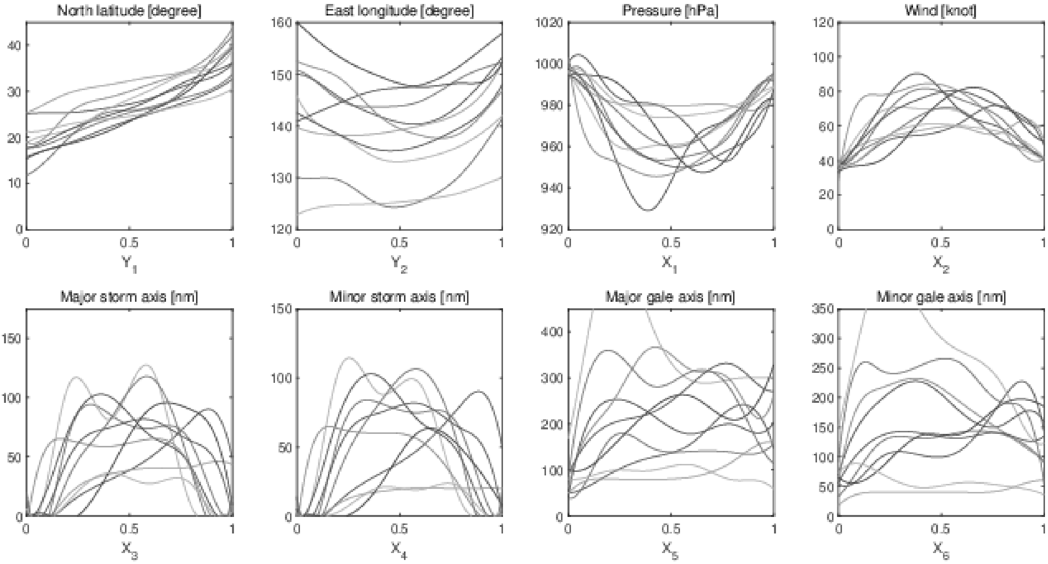

Figure 4: Functional data sets obtained by smoothing the data given in Figure 3.

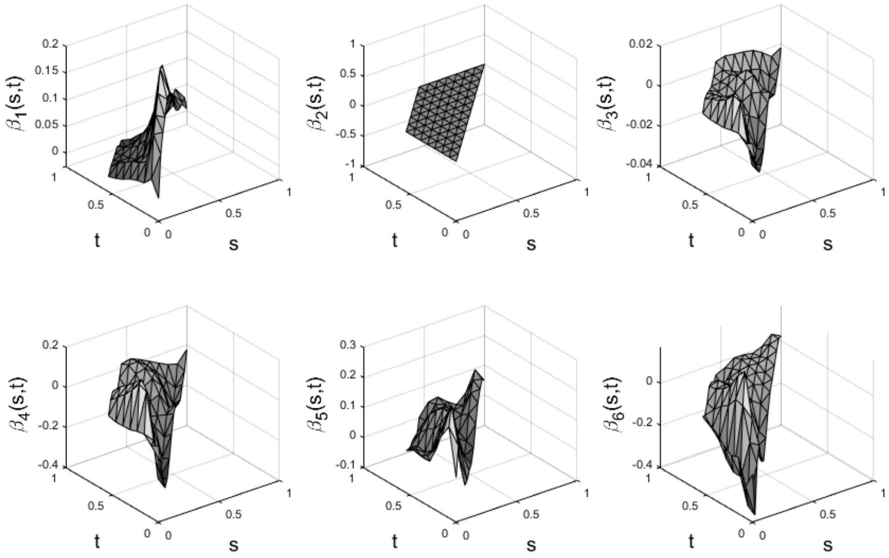

Figure 5: Estimated coefficient functions when the response is the north latitude (Y_1).

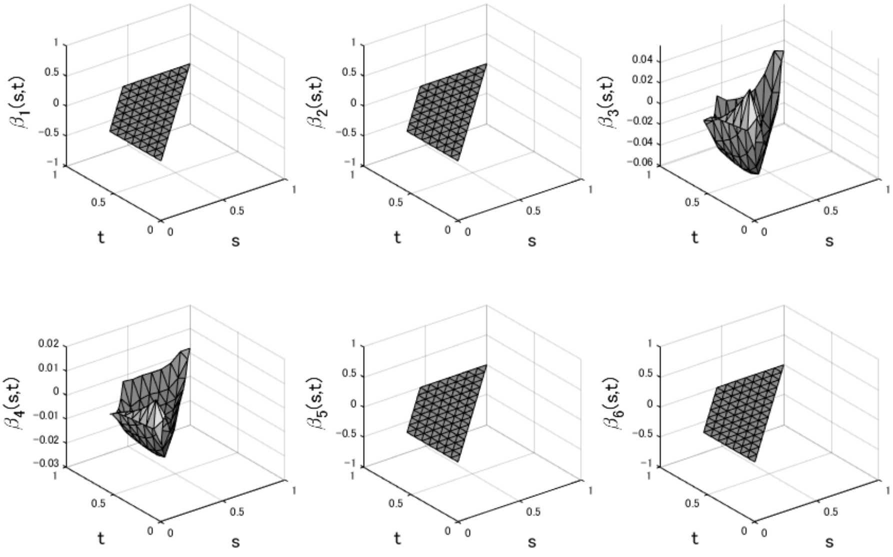

Figure 6: Estimated coefficient functions when the response is the east longitude (Y_2).