Published in final edited form as:

Commun Stat Theory Methods. 2017 ; 46(15): 7703–7716. doi:10.1080/03610926.2016.1161796.

Automatic Structure Discovery for Varying-coefficient Partially Linear Models

Guangren Yang [Lecturer],

Department of Statistics, School of Economics, Jinan University, Guangzhou, P.R. China.
tygr@jnu.edu.cn

Yanqing Sun [Professor], and

Department of Mathematics and Statistics, University of North Carolina at Charlotte.
yasun@uncc.edu

Xia Cui* [Associate professor]

School of Economics and Statistics, Guangzhou University, Guangzhou, P.R. China.
cuixia@gzhu.edu.cn

Abstract

Varying-coefficient partially linear models provide a useful tools for modeling of covariate effects on the response variable in regression. One key question in varying-coefficient partially linear models is the choice of model structure, that is, how to decide which covariates have linear effect and which have nonlinear effect. In this article, we propose a profile method for identifying the covariates with linear effect or nonlinear effect. Our proposed method is a penalized regression approach based on group minimax concave penalty. Under suitable conditions, we show that the proposed method can correctly determine which covariates have a linear effect and which do not with high probability. The convergence rate of the linear estimator is established as well as the asymptotical normality. The performance of the proposed method is evaluated through a simulation study which supports our theoretical results.

Keywords

B-spline; group MCP; varying-coefficient partially linear models; coordinate descent algorithm

AMS 2001 subject classification:

62J05; 62G08; 62G20

1 Introduction

Suppose we have a random sample $(y_i, u_i, x_{i1}, \dots, x_{ip}), 1 \leq i \leq n$, where y_i is the response variable, u_i is a one-dimensional covariate and $(x_{i1}, \dots, x_{ip})$ is a p -dimensional covariate vector. Consider the varying-coefficient partially linear model

*Address for correspondence.

$$y_i = \sum_{j \in S_1} \beta_{0j} x_{ij} + \sum_{j \in S_2} \beta_{1j} u_i x_{ij} + \sum_{j \in S_3} g_j(u_i) x_{ij} + \varepsilon_i, 1 \leq i \leq n, \quad (1.1)$$

where S_1 , S_2 and S_3 are mutually exclusive and complementary subsets of $\{1, \dots, p\}$, $\{\beta_{0j} : j \in S_1\}$ are constant regression coefficients of the covariates with indices in S_1 , $\{\beta_{1j} u : j \in S_2\}$ are linear regression coefficients of the covariates with indices in S_2 , and $\{g_j(u) : j \in S_3\}$ are unknown functions. In this model, the mean response is linearly related to the covariates in S_1 and S_2 , while its relation with the remaining covariates is not specified up to any finite number of parameters. This model combines the flexibility of varying-coefficient regression and the parsimony of linear regression. When the relation between y_i and $\{x_{ij} : j \in S_1\}$ or $\{x_{ij} u_i : j \in S_2\}$ is of main interest and can be approximated by a linear function, it offers more interpretability than a purely varying-coefficient model.

The varying-coefficient models are not stimulated by the desire of purely mathematical extension, rather they come from the need in practice. In many scientific areas where statistics is needed, there are some commonly used traditional parametric models found by the people in the area in the light of their experience. The varying coefficient models are introduced by Cleveland, Grosse and Shyu (1991) to extend the applications of local regression techniques from one-dimensional to multi-dimensional setting. An extension of the local regression was given by Hastie and Tibshirani (1993). There are three approaches to estimate the nonparametric function in model (1.1). One is kernel-local polynomial smoothing, see Wu et al. (1988), Hoover et al. (1998), Fan and Zhang (1999), Kauermann and Tutz (1999). One is polynomial spline, see Huang et al. (2002, 2004) and Huang and Shen (2004). The last one is smoothing spline, see Hastie and Tibshirani (1993), Hoover et al. (1998) and Chiang et al. (2001). Fan and Huang (2005) have more deeply investigated model (1.1). They proposed a profile least-squares technique to estimate parametric part, and established the asymptotic normality of the estimator.

More recently, variable selection methods using penalty functions in nonparametric or semiparametric settings have been developed. For generalized varying-coefficient partial linear models, Li and Liang (2008) used penalization to select the significant predictors in the parametric components while the nonparametric components were selected by hypothesis testing. Note that this work is for the fixed p case. Lam and Fan (2008) studied varying-coefficient partially linear models with a diverging number of parametric components but they did not investigate the variable selection problem. For the ordinary varying-coefficients models with quadratic loss function, Wang and Xia (2008) proposed a group penalization method in the fixed p case, and Wei, Huang, and Li (2011) recently extended this work to the case of diverging p . Huang, Wei, and Ma (2012) proposes a semiparametric regression pursuit method for indentifying the covariates with a linear effect by using a group minimax concave penalty in the semiparametric partially linear models and showed that the proposed approach is model-pursuit consistent, meaning that it can correctly determine which covariates have a linear effect and which do not with high probability.

In this article, we propose a profile approach for indentifying the covariates with a linear effect by using a group minimax concave penalty in the varying-coefficient partially linear models (1.1). We show that the group MCP possesses the asymptotic oracle property, meaning that the group MCP estimator is equal to the oracle least squares estimator assuming the model is known with probability converging to one. The convergence rate of the linear estimator is established as well as the asymptotical normality. We also derive a group coordinate descent algorithm for calculating the group MCP estimators for this model. Our simulation studies show that the proposed method is effective in fitting (1.1) and indicate that the group MCP method works well for this model.

The paper is organized as follows. In section 2, we describe the procedure for selection and estimation using the group MCP based on the basis expansion for the regression functions. Section 3 introduces the results on the selection consistency of the group MCP in high-dimensional setting and the parametric part can obtain $\sqrt{n}$ convergence rate and asymptotical normality. Section 4 provides a simulation study is used to illustrate the proposed methods. Concluding remarks are given in Section 5. Proofs are given in Appendix.

2 Method

The varying-coefficient partially linear model (1.1) can be embedded into the varying-coefficient model (Hastie and Tibshirani (1993)),

$$y_i = \alpha_1(u_i)x_{i1} + \cdots + \alpha_p(u_i)x_{ip} + \varepsilon_i. \quad (2.1)$$

If some of the α_j 's are linear, then (2.1) becomes the varying-coefficient partially linear model (1.1). The problem is that of determining which $\alpha_j(u)$'s have a constant and linear forms and which do not. Without loss of generality, suppose that $u_1, \dots, u_n$ are from a distribution with support $[a, b]$, where $a < b$ are finite constants. For this purpose, we write $\alpha_j(u)$ as

$$\alpha_j(u) = \beta_{0j} + \beta_{1j}u + g_j(u). \quad (2.2)$$

Consider a truncated series expansion for approximating $g_j(u)$,

$$g_{nj}(u) = \sum_{k=1}^{m_n} \theta_{jk} \psi_k(u), \quad j = 1, \dots, p. \quad (2.3)$$

where $\{\psi_k(u), k = 1, \dots, m_n\}$ is a set of m_n certain basis functions, m_n is allowed to increase with n and here we can take the B-spline basis functions. According to the classical book of Schumaker (1981), B-spline has a good property since the B-splines are nonnegative and sum to 1 at any point. Furthermore, there doesn't exist identification problem when nonparametric function is expanded by using B-splines. If $\theta_{jk} = 0, 1 \leq k \leq m_n$, then $\alpha_j(u)$'s is

of linear structure. Therefore, with this formulation, the problem is to determine which groups $(\theta_{jk}; 1 \leq k \leq m_n), j \in S_3$, are zero.

Let $\beta_0 = (\beta_{01}, \dots, \beta_{0p})'$, $\beta_1 = (\beta_{11}, \dots, \beta_{1p})'$ and $\theta = (\theta'_1, \dots, \theta'_p)'$, where $\theta_j = (\theta_{j1}, \dots, \theta_{jm_n})'$, $1 \leq j \leq p$. Define the penalized least squares criterion

$$L(\beta_0, \beta_1, \theta; \lambda, \gamma) = \frac{1}{2n} \sum_{i=1}^n \|y_i - \sum_{j=1}^p \beta_{0j} x_{ij} - \sum_{j=1}^p \beta_{1j} u_i x_{ij} - \sum_{j=1}^p \sum_{k=1}^{m_n} \theta_{jk} \psi_k(u_i) x_{ij}\|^2 \quad (2.4) \\ + \sum_{j=1}^p \rho(\|\theta_j\|_{A_j}; \sqrt{m_n} \lambda, \gamma),$$

where $\rho(\cdot; \cdot, \cdot)$ is a penalty function depending on the penalty parameter $\lambda \geq 0$ and a regularization parameter γ . The norm $\|\theta_j\|_{A_j} = (\theta'_j A_j \theta_j)^{1/2}$ for a given positive definite matrix A_j . In theory, any positive definite matrix can be used as A_j since $\|\theta_j\|_{A_j} = 0$ if and only if $\theta_j = 0$. However, it is important to choose a suitable A_j to make the amount of penalization comparable across the groups and to facilitate the computation. We use the minimax concave penalty (MCP) and detail discussions on the MCP can be found in Zhang (2010). The MCP provides the convexity of the penalized loss in sparse regions to the greatest extent given certain thresholds for variable selection and unbiasedness. Its output is a continuous piecewise linear path encompassing from the origin for infinite penalty to a least squares solution for zero penalty. Zhang (2010) proves that the MCP attains certain minimax convergence rates in probability for the estimation of regression coefficients.

We use the minimax concave penalty or MCP, introduced by Zhang (2010), given by

$$\rho(t; \lambda, \gamma) = \lambda \int_0^t \left(1 - \frac{x}{\gamma \lambda}\right)_+ dx, \quad t \geq 0, \quad (2.5)$$

where γ is a parameter that controls the concavity of ρ and λ is the penalty parameter. Here $x_+ = x \mathbb{I}(x \geq 0)$. We require $\lambda \geq 0$ and $\gamma > 1$. The term MCP comes from the fact that it minimizes the maximum concavity measure defined at (2.2) of Zhang (2010), subject to conditions on unbiasedness and selection features.

The penalty in (2.4) is a composite of the penalty function $\rho(\cdot; \lambda, \gamma)$ and a weighted l_2 -norm of θ_j . The $\rho(\cdot; \lambda, \gamma)$ is a penalty for individual variable selection. When applied to a norm of θ_j , it selects the coefficients in θ_j as a group. This is desirable, since the nonlinear components are represented by the coefficients in the θ_j 's as groups. Based on (2.4), it is natural to call it the group minimax concave penalty (gMCP).

For simplicity, let $\beta = (\beta'_0, \beta'_1)'$, $y = (y_1, \dots, y_n)'$, $X_i = (x_{i1}, \dots, x_{ip}, u_i x_{i1}, \dots, u_i x_{ip})$, $Z_i = (x_{i1}, \dots, x_{ip}) \otimes (\psi_1(u_i), \dots, \psi_{m_n}(u_i))$, $i = 1, \dots, n$, where $\otimes$ is tensor product. Let $X = (X'_1, \dots, X'_n)$ and $Z = (Z'_1, \dots, Z'_n)$, (2.4) can be rewritten as

$$L(\beta, \theta; \lambda, \gamma) = \frac{1}{2n} \|y - X\beta - Z\theta\|^2 + \sum_{j=1}^p \rho(\|\theta_j\|_{A_j}; \sqrt{m_n} \lambda, \gamma). \quad (2.6)$$

For a given (λ, γ) , the penalized least squares solution is

$$(\hat{\beta}_n, \hat{\theta}_n) = \arg \min_{\beta, \theta} L(\beta, \theta; \lambda, \gamma).$$

To compute $(\hat{\beta}_n, \hat{\theta}_n)$, we can use a penalized profile least squares approach. For any given θ_n , the $\hat{\beta}$ that minimizes L necessarily satisfies

$$X'(y - X\beta - Z\theta_n) = 0.$$

Thus $\hat{\beta} = (X'X)^{-1}X'(y - Z\theta_n)$. Let $Q = I - P_X$, where $P_X = X(X'X)^{-1}X'$ is the projection matrix onto the column space of X . The profile objective function of θ_n is

$$L(\theta_n; \lambda, \gamma) = \frac{1}{2n} \|Q(y - Z\theta_n)\|^2 + \sum_{j=1}^p \rho(\|\theta_{nj}\|_{A_j}; \sqrt{m_n} \lambda, \gamma). \quad (2.7)$$

As noted above, any positive definite matrix can be used for A_j . Here we use $A_j = n^{-1}U_j'QU_j$, where Q is an orthonormal matrix.. The rationale for this choice is based on the following considerations. The covariate matrix for group j is QU_j in the profile objective function (2.7). Therefore, this choice of A_j standardizes the covariate matrices associated with θ'_{nj} s and makes the amount of penalization comparable across the groups comparable.

Second, it leads to an explicit expression in the update steps in the group coordinate algorithm describes below; this facilitates the implementation of the algorithm, since computation in each update step can be carried out using explicit expressions. For any given (λ, γ) , the penalized profile least squares solution is $\hat{\theta}_n = \arg \min_{\theta_n} L(\theta_n; \lambda, \gamma)$.

Here, we will follow Huang (2012). The set of indices of the covariates that are estimated to have the linear structure in the regression model (1.1) is $S_1 \hat{\cup} S_2 \equiv \{j : \|\hat{\theta}_{nj}\|_2 = 0\}$. Thus,

$$\hat{g}_{nj}(u) = 0, j \in S_1 \hat{\cup} S_2 \quad \text{and} \quad \hat{g}_{nj}(u) = \sum_{k=1}^{m_n} \hat{\theta}_{jk} \psi_k(u), j \notin S_1 \hat{\cup} S_2.$$

Let $\hat{X}_{(1)} = (x_j, ux_j), j \in S_1 \hat{\cup} S_2$, $\hat{Z}_{(2)} = (Z_j : j \notin S_1 \hat{\cup} S_2)$ and $\hat{\theta}_{n(2)} = (\hat{\theta}'_{nj} : j \notin S_1 \hat{\cup} S_2)'$. We have $\hat{\beta}_n = (X'X)^{-1}X'(y - \hat{Z}_{(2)}\hat{\theta}_{n(2)})$. The estimator of the coefficients of the linear components is $\hat{\beta}_{n1} = (\hat{\beta}'_j : j \in S_1 \hat{\cup} S_2)'$. Let

$$\hat{\alpha}_{nj}(u) = (1, u)\hat{\beta}_j + \hat{g}_j(u), \quad j \notin S_1 \cup S_2.$$

where $\hat{\beta}_j$ corresponds to the coefficients of the linear components, i.e. (x_j, ux_j) . Let us write $\hat{\alpha}_n(u) = (\hat{\alpha}_1(u), \dots, \hat{\alpha}_p(u))'$. Then the estimator of the coefficient vector of the linear components can also be written as

$$\hat{\beta}_{n1} = (\hat{X}'_{(1)}\hat{X}_{(1)})^{-1}\hat{X}'_{(1)}\left(y - \sum_{j \notin S_1 \cup S_2} \hat{\alpha}_{nj}(u)x_j\right).$$

3 Theoretical properties

We present results on the model consistency, rate of convergence and asymptotic normality of the proposed profile estimator of varying-coefficient partially linear models. In particular, our model consistency result shows that the proposed method can correctly determine the linear and nonlinear components in the varying-coefficient partially linear models with high probability.

Denote the underlying regression components by $\alpha_{0j}(u)$ and write

$$\alpha_{0j}(u) = \beta_{0j} + \beta_{1j}u + g_{0j}(u).$$

Suppose the series expansion for approximating $g_{0j}(u)$ is

$$g_{0j}(u) = \sum_{k=1}^{m_n} \theta_{0jk} \psi_k(u).$$

Let $\theta_{0jn} = (\theta_{0j1}, \dots, \theta_{0jm_n})'$, and let $\|g\|_2 = [Eg^2(u)]^{1/2}$ for any square integrable function g on $[a, b]$. We have $S_1 \cup S_2 = \{j: \|\alpha_{0j}(u)\|_2 = 0\}$ and $\|\theta_{nj}\|_2 = 0$ for $j \in S_1 \cup S_2$. Let $\theta_{0n} = (\theta'_{0n1}, \dots, \theta'_{0np})'$.

Let $q = |S_1 \cup S_2|$ be the cardinality of $S_1 \cup S_2$, the number of linear components in the regression model. Take

$$\tilde{\theta}_n = \arg \min_{\theta} \left\{ \frac{1}{2n} \|Q(y - Z\theta_n)\|^2 : \theta_{nj} = 0, \quad j \in S_1 \cup S_2 \right\}. \quad (3.1)$$

This is the oracle estimator of θ_{0n} that takes the identity of the linear components as known.

Analogous to the estimates defined at the end of Section 2, write the oracle estimators as

$$\tilde{g}_{nj}(u) = 0, \quad j \in S_1 \cup S_2 \quad \text{and} \quad \tilde{g}_{nj}(u) = \sum_{k=1}^{m_n} \tilde{\theta}_{jk} \psi_k(u), \quad j \notin S_1 \cup S_2.$$

Let $X_{(1)} = (x_j, j \in S_1 \cup S_2)$, $X_{(2)} = (x_j : j \in S_3)$, $\tilde{\theta}_{n(2)} = (\tilde{\theta}'_{nj}, j \in S_3)'$, and

$$\tilde{\alpha}_{nj}(u) = \tilde{\beta}_{0j} + \tilde{\beta}_{1j}u + \tilde{g}_{nj}(u), \quad j \in S_3.$$

The oracle estimator of the coefficients of the linear components is

$$\tilde{\beta}_{n1} = (X'_{(1)}X_{(1)})^{-1}X'_{(1)}\left(y - \sum_{j \in S_3} \tilde{\alpha}_{nj}(u)x_{ij}\right).$$

Without loss of generality, suppose that $S_1 \cup S_2 = \{1, \dots, q\}$. Write $\tilde{\theta}_n = (0'_{qm_n}, \tilde{\theta}'_{n(2)})'$, where 0_{qm_n} is a (qm_n) -dimensional vector of zeros and

$$\tilde{\theta}_{n(2)} = (Z'_{(2)}QZ_{(2)})^{-1}Z'_{(2)}Qy. \quad (3.2)$$

where $Z_{(2)}$ is the design matrix related to $\theta_{n(2)}$. Let $\theta_* = \min_{j \in S_1 \cup S_2} \|\theta_{0nj}\|$, the smallest norm of the coefficients in the spline expansions of the nonlinear components.

Let k be a non-negative integer, and let $\alpha \in (0, 1]$ be such that $d = k + \alpha > 0.5$. Let $\mathcal{G}$ be the class of functions g on $[0, 1]$ whose k th derivative $g^{(k)}$ exists and satisfies a Lipschitz condition of order α :

$$|g^{(k)}(s) - g^{(k)}(t)| \leq C |s - t|^\alpha \quad \text{for } s, t \in [a, b].$$

Then there exists a positive constant C_1 such that

$$\|g_{0j}(u) - g_{nj}(u)\| = O(m_n^{-2r-2\alpha}).$$

We make the following assumptions.

(A1) p and q are fixed and $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed with $E\varepsilon_i = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. Furthermore, $P(|\varepsilon_i| > x) \leq K \exp(-Cx^2)$, $i = 1, \dots, n$, for all $x > 0$ for some constants C and K .

(A2) The covariate vector u has a continuous density and there exist constants C_1 and C_2 such that the density function $\alpha_j(\cdot)$ of u satisfies $0 < C_1 \leq \alpha_j(u) \leq C_2 < \infty$ on $[a, b]$ for every $1 \leq j \leq p$.

(A3) There exists a positive constant M such that $|x_{ij}| \leq M, i = 1, \dots, n, j = 1, \dots, p$.

Theorem 1 Suppose that $m_n = O(n^{1/(2d+1)})$, $1/\sqrt{m_n}\gamma$ is less than the smallest eigenvalue of $Z'QZ/n$, and

$$\frac{1}{m_n^{(2d-1)/2}(\theta_* - \lambda\gamma)} + \frac{1}{\lambda\sqrt{n}} \rightarrow 0. \quad (3.3)$$

Then under (A1)—(A3), $P(\hat{\theta}_n \neq \tilde{\theta}_n) \rightarrow 0$. Consequently, $P(S_1 \hat{\cup} S_2 = S_1 \cup S_2) \rightarrow 1$, and

$$P(\hat{\beta}_{n1} = \tilde{\beta}_{n1}) \rightarrow 1, \quad \text{and} \quad P(\|\hat{\alpha}_{nj}(u) - \tilde{\alpha}_{nj}(u)\|_2 = 0, j \in S_3) \rightarrow 1.$$

Therefore, under the conditions of Theorem 1, the proposed estimator can correctly distinguish constant, linear and nonlinear components with high probability. Furthermore, the proposed estimator has the oracle property in the sense that it is the same as the oracle estimator assuming the identity of the constant, linear and nonlinear components were known, except on an event with probability tending to zero. The main extra condition here is (3.3), which requires both $\lambda = O(n^{-1/2})$ and $\theta_* > \gamma\lambda + a_n m_n^{-(2d-1)/2}$ for some $a_n \rightarrow \infty$. The first part of this requirement ensures that the bias resulting from the penalty is so small that it does not interfere with selection, and the second part requires that the smallest norm θ_* of the coefficients in the spline expansions of the (nonzero) nonlinear components be larger than the penalty level plus a term due to the spline approximation error.

Theorem 2 Under (A1)—(A3) hold, we have

$$\sum_{j=1}^p \|\hat{\alpha}_{nj}(u) - \alpha_{0j}(u)\|_2^2 \leq O_p(m_n/n) + O(1/m_n^{2d}) + O(m_n\lambda^2).$$

This theorem gives rate of convergence of the proposed estimator of nonparametric regression coefficient under the semivary-coefficient model (1.1). In particular, if we assume that each nonparametric function component in (1.1) is second-order differentiable ($d=2$) and take $m_n = O(n^{1/5})$ and $\lambda = n^{-1/2+\delta}$ for a small $\delta > 0$, then $\sum_{j=1}^p \|\hat{\alpha}_{nj} - \alpha_{0j}\|_2^2 = O_p(n^{-4/5})$, which is the optimal rate of convergence in nonparametric function.

We now consider the asymptotic distribution of $\hat{\beta}_{n1}$. According to Theorem 3.2. of Fan and Huang (2005), we can obtain Theorem 3 as follows.

Theorem 3 If the conditions (A1)—(A3) are satisfied, and Theorems 1 and 2 also hold, then

$$\sqrt{n}(\hat{\beta}_{n1} - \beta_{(1)}) \xrightarrow{d} N(0, \Sigma)$$

where $\Sigma = (E(X_{(1)}X_{(1)}^T) - E[E(X_{(1)}Z_{(2)}^T)\{E(Z_{(2)}Z_{(2)}^T)\}^{-1}E(Z_{(2)}X_{(1)}^T)])^{-1}\sigma^2$ and $\beta_{(1)} = (\beta_j: j \in S_1 \cup S_2)'$.

The limit distribution of Theorem 3 is the same as that of the oracle estimator $\tilde{\beta}_{n1}$.

4 Numerical studies

In this section, We conduct simulation studies to evaluate the finite-sample performance of the group MCP and group Lasso in (1.1).

We use the group coordinate descent algorithm for computing the group MCP solutions for the varying coefficient model (1.1). For the group Lasso in linear regression, a group coordinate descent algorithm was described in Yuan and Lin (2006). It is a natural extension of standard coordinate descent, see for example, Fu (1998). Meier et al. (2008) also used a group coordinate descent for group selection in high-dimensional logistic regression. Breheny and Huang (2011) considered coordinate descent algorithm for nonconvex penalized regression. The algorithm we used here takes the same form as those described in Huang et al. (2012), so we refer to that paper for a detailed description. However, we need to take into account the specific nature of the model (1.1).

We used simulation to evaluate the finite sample performance of the proposed method. Two samples were considered in the simulation. In each of the simulated models, two sample sizes ($n=400,800$) were conducted. Because our main interest is to distinguish constant, linear and nonlinear effects. We consider the model

$$y_i = \sum_{j \in S_1} \beta_{0j} x_{ij} + \sum_{j \in S_2} \beta_{1j} u_i x_{ij} + \sum_{j \in S_3} \alpha_j(u_i) x_{ij} + \varepsilon_i.$$

We simulated as follows: $u_i \stackrel{i.i.d}{\sim} U[0, 1]$, $x_{ik} \stackrel{i.i.d}{\sim} N(0, 1)$ and $\varepsilon_i \stackrel{i.i.d}{\sim} N(0, 1)$ where $i = 1, \dots, n$, $k = 1, \dots, p$. The coefficient functions are

$$\begin{aligned} \alpha_1(u) &= 3, & \alpha_1(u) &= 1.2, & \alpha_3(u) &= 1.5 \sin(2\pi u) / (2 - \sin(2\pi u)), \\ \alpha_4(u) &= 2(\sin(2\pi u))^3, & \alpha_5(u) &= (3u - 1)^2, \\ \alpha_6(u) &= 2.2 \cos(2\pi u) / (2 - \cos(2\pi u)). \end{aligned}$$

Example 1. Let $p = 4$ and consider the model

$$y = 3x_1 + 1.2ux_2 + \alpha_3(u)x_3 + \alpha_4(u)x_4 + \varepsilon.$$

Here the first variable is constant effect, the second variable is linear effect and the last two have nonlinear effect.

Example 2. Let $p = 6$ and consider the model

$$y = 3x_1 + 1.2ux_2 + \alpha_3(u)x_3 + \alpha_4(u)x_4 + \alpha_5(u)x_5 + \alpha_6(u)x_6 + \varepsilon.$$

Here the first variable is constant effect, the second variable is linear effect and the last four have nonlinear effect.

We use the cubic B-spline basis functions for each α_k . In our application, we apply the bayesian information criterion (BIC) (Schwarz (1978)) to select (λ, m_n) for the group MCP. The BIC is given

$$BIC(\lambda, m_n) = \log(RSS_{\lambda, m_n}) + \log n \frac{df_{\lambda, m_n}}{n},$$

where RSS is the residual sum of squares, df is the number of selected variables for a given (λ, m_n) . We choose m_n from an increasing sequence of 10 knots; for any given value of m_n , we choose λ from a sequence of 100 values, starting from $\lambda_{\max}$ to $0.01\lambda_{\max}$. This $\lambda_{\max}$ is the smallest penalty value that forces all the estimated coefficients to be zero.

The simulation results based on 100 replications are presented in Tables 1–3. The columns in Table 1 are the average number of nonlinear components being selected (NL), the average model error (ER), the percentage of occasions in which the correct nonlinear components were included in the selected model (IN%), and the percentage of occasions in which the exactly nonlinear components were selected (CS%) in the final model. Enclosed in parentheses are the corresponding standard errors. Tables 2 and 3 include the number of times each component estimated as a nonlinear function. Enclosed in parentheses are the corresponding standard errors.

Table 1 shows that the proposed method with the group MCP performed better than the proposed method with the group Lasso in terms of the percentage of occasions on which the correct nonlinear components were included in the selected model (IN%) and the percentage of occasions in which the exactly nonlinear components were selected (CS%) in the final model. For instance, in Example 1 with $n=400$, the percentage of correct selection NL was 14.44 with the group MCP and 7.14 with the group Lasso. Also, when the sample size was 800, the percentage of inclusion of all the nonlinear components approximate true nonlinear components. From table 1, we show that the group MCP was more accurate in distinguishing the linear functions from the nonlinear functions than the group Lasso. In Tables 2 and 3, we examine the performance of the proposed method for estimating the linear and nonlinear components in the simulated models. In general, the proposed method with the group MCP has smaller mean square errors. Overall, the proposed method with the group MCP was effective in distinguishing the linear components from the nonlinear ones in the simulation models.

5 Conclusion

In this paper, we proposed a varying-coefficient partially linear regression profile method for distinguishing constant and linear from nonlinear components based on group MCP method. This approach determines the parametric and nonparametric components in varying-coefficient partially linear model adaptively based on the data. Our proposed method is fundamentally different from the standard varying-coefficient inference approach where the parametric and nonparametric components in a model are pre-specified. We showed that our method has the asymptotic oracle properties under the group MCP method, meaning that it is the same as the standard varying-coefficient partially linear regression estimator assuming the model structure is known. The asymptotic rates of the penalty parameters required for our theoretical results are derived. However, as in many recent studies, it is not clear whether the penalty parameters selected using the BIC or other procedures can match the asymptotic rates. This is an important and challenging problem that requires further investigation, but is beyond the scope of the current paper. Our simulation study indicates that the proposed method works well in finite sample situations.

Acknowledgement:

We thank Professor Jian Huang for providing this idea of this article and for many help. Yang's research was supported by the Fundamental Research Funds for the Central Universities (No. 15JNQM019 and 21615452), the National Statistical Scientific Research Center Projects (No. 2015LD02) and the China Scholarship Council (No. 201506785010). Work of Cui is supported by a National Nature Science Foundation of China grant (No. 11471086), the Excellent Young Teachers Program of Department of Education of Guangdong Province (No. Yq201404) and Guangzhou University Project of Humanities and Social Sciences (No. 201503XSTD).

Appendix: Proofs

Lemma 1 If

$$\frac{(p-q)^{1/2}}{m_n^{(2d-1)/2}(\theta_* - \gamma\lambda)} \rightarrow 0,$$

then

$$P\left(\max_{j \notin S_1 \cup S_2} \|\tilde{\theta}_{nj} - \theta_{nj}\| > \theta_* - \gamma\lambda\right) \leq O(1) \frac{(p-q)m_n}{\sqrt{n}(\theta_* - \gamma\lambda)}. \quad (\text{A.1})$$

Proof of Lemma 1. Let T_{nj} be an $m_n \times (p-q)m_n$ matrix with the form

$$T_j = (0_{m_n}, \dots, 0_{m_n}, I_{m_n}, 0_{m_n}, \dots, 0_{m_n})$$

where 0_{m_n} is an $m_n \times m_n$ matrix of zeros and I_{m_n} is an $m_n \times m_n$ identity matrix in the j th block. By the triangle inequality,

$$\|\tilde{\theta}_{nj} - \theta_{nj}\|_2 \leq \|T_{nj}C_{(2)}^{-1}Z'_{(2)}Q\varepsilon_n\|_2 + \|T_{nj}C_{(2)}^{-1}Z'_{(2)}Q\delta_n\|_2. \quad (\text{A.2})$$

Let C be a generic constant independent of n . For the first term on the right-hand side, we have

$$\begin{aligned} E \max_{j \notin S_1 \cup S_2} \|T_{nj}C_{(2)}^{-1}Z'_{(2)}Q\varepsilon_n\|_2 &\leq n^{-1}\rho_{n1}^{-1}E\|Z'_{(2)}Q\varepsilon_n\|_2 \\ &= n^{-1/2}\rho_{n1}^{-1}E\|n^{-1/2}Z'_{(2)}Q\varepsilon_n\|_2 \\ &= n^{-1/2}\rho_{n1}^{-1}m_n^{-1/2}((p-q)m_n)^{1/2} \\ &= O(1)(p-q)n^{-1/2}m_n. \end{aligned}$$

Thus

$$P\left(\max_{j \notin S_1 \cup S_2} \|T_{nj}C_{(2)}^{-1}Z'_{(2)}Q\varepsilon_n\| \geq (\theta_* - \gamma)/2\right) \leq \frac{O(1)(p-q)m_n}{\sqrt{n}(\theta_* - \gamma\lambda)}.$$

By the approximation properties of splines to a smooth function, we have $n^{-1}\|\delta_n\|_2^2 = O_p((p-q)m_n^{-2d})$ and, for the second term,

$$\begin{aligned} \max_{j \notin S_1 \cup S_2} \|T_{nj}C_{(2)}^{-1}Z'_{(2)}Q\delta_n\|_2 &\leq \|nC_{(n)}^{-1}\|_2 \cdot \|n^{-1}Z'_{(2)}Z_{(2)}\|_2^{1/2} \cdot \|n^{-1/2}\delta_n\|_2 \\ &= O_p(1)\rho_{n1}^{-1}\rho_{n2}^{-1/2}(p-q)^{1/2}m_n^{-d} \\ &= O_p(1)(p-q)^{1/2}m_n^{-(2d-1)/2}. \end{aligned}$$

Therefore, when

$$\frac{(p-q)m_n}{\sqrt{n}(\theta_* - \gamma\lambda)} \rightarrow 0,$$

then (A.2) holds.

Lemma 2 If

$$\frac{1}{\lambda m_n^{(2d+1)/2}} \rightarrow 0, \quad (40)$$

then

$$P(n^{-1} \max_{j \in S_1 \cup S_2} \|Z_j' H(\varepsilon_n + \delta_n)\| > \lambda) \leq O(1) \frac{\{\log\{(q \vee 1)m_n\}\}^{1/2}}{\lambda \sqrt{n}}. \quad (\text{A.3})$$

Proof of Lemma 2. Write

$$n^{-1} Z_j' H(\varepsilon_n + \delta_n) = n^{-1} Z_j' H_n \varepsilon_n + n^{-1} Z_j' H_n \delta_n. \quad (\text{A.4})$$

By Lemma 2 of Huang et al. (2010)

$$E \left(\max_{j \in S_1 \cup S_2} \|n^{-1/2} Z_j' H_n \varepsilon_n\|_2 \right) \leq O(1) \{\log((p - |S_1 \cup S_2|)m_n)\}^{1/2}. \quad (\text{A.5})$$

Therefore,

$$P(n^{-1} \max_{j \in S_1 \cup S_2} \|Z_j' H_n \varepsilon_n\|_2 > \lambda/2) \leq O(1) \frac{\{\log(qm_n)\}^{1/2}}{\lambda \sqrt{n}}. \quad (\text{A.6})$$

For the second term on the right hand side of (A.4)

$$\begin{aligned} n^{-1} \max_{j \in S_1 \cup S_2} \|Z_j' H_n \delta_n\|_2 &\leq n^{-1/2} \max_{j \notin S_3} \|n^{-1} Z_j' Z_j\|_2^{1/2} \cdot \|H_n\|_2 \cdot \|\delta_n\|_2 \\ &= O(1) \rho_{n2}^{1/2} (p-q)^{1/2} m_n^{-d} \\ &= O(1) (p-q)^{1/2} m_n^{-(2d+1)/2}. \end{aligned}$$

Therefore, when

$$\frac{1}{\lambda m_n^{(2d+1)/2}} \rightarrow 0,$$

then

$$P(n^{-1} \max_{j \in S_1 \cup S_2} \|Z_j' H(\varepsilon_n + \delta_n)\| > \lambda) \leq O(1) \frac{\{\log\{(q \vee 1)m_n\}\}^{1/2}}{\lambda \sqrt{n}}.$$

Proof of Theorem 1 Since $1/\sqrt{m_n}\lambda$ is less than the smallest eigenvalue of $Z'QZ/n$, $L(\cdot; \lambda, \gamma)$ in (2.7) is a convex function. By the Karush-Kuhn-Tucker conditions, a necessary and sufficient condition for $\hat{\theta}$ is

$$\begin{cases} Z'_j Q(y - Z\hat{\theta}_n) = n\dot{\rho}(\|\hat{\theta}_n\|; \lambda), & \|\hat{\theta}_{nj}\|_2 \neq 0, \\ \|Z'_j Q(y - Z\hat{\theta}_n)\|_2 \leq n\lambda, & \|\hat{\theta}_{nj}\|_2 = 0, \end{cases} \quad (\text{A.7})$$

For $j \notin S_1 \cup S_2$, if $\|\tilde{\theta}_{nj}\| \geq \gamma\lambda$, then $\dot{\rho}(\|\tilde{\theta}_{nj}\|; \lambda) = 0$. Thus $\tilde{\theta}_n$ satisfies (A.7) if also $\|Z'_j Q(y - Z\tilde{\theta}_n)\|_2 \leq n\lambda$ for $j \notin S_3$. Therefore, $\hat{\theta}_n = \tilde{\theta}_n$ is the intersection of the events

$$\Omega_1(\lambda) = \left\{ \min_{j \in S_3} \|\tilde{\theta}_{nj}\| \geq \gamma\lambda \right\} \quad \text{and} \quad \Omega_2(\lambda) = \left\{ \min_{j \notin S_3} \|Z'_j Q(y - Z\tilde{\theta}_n)\| \leq n\lambda \right\}.$$

Let $g_{0j}(x_j) = (g_{0j}(x_{1j}), \dots, g_{0j}(x_{nj}))'$ and $\delta_n = \sum_{j \in S_3} g_{0j}(x_j) - Z_{(2)}\theta_{n(2)}$. Let $C_{(2)} = Z'_{(2)}QZ_{(2)}$ and $H = Q - QZ_{(2)}(Z'_{(2)}QZ_{(2)})^{-1}Z'_{(2)}Q$. By (3.2),

$$\begin{aligned} \tilde{\theta}_{n(2)} - \theta_{n(2)} &= C_{(2)}^{-1}Z'_{(2)}Q(\varepsilon_n + \delta_n), \\ Z'_j Q(y - Z_{(2)}\tilde{\theta}_{n(2)}) &= Z'_j H(\varepsilon_n + \delta_n), \end{aligned}$$

Recall that $\theta_* = \min_{j \in S_3} \|\theta_{nj}\|$. If $\|\tilde{\theta}_{nj} - \theta_{nj}\| \leq \theta_* - \gamma\lambda$, then $\min_{j \in S_3} \|\tilde{\theta}_{nj}\| \geq \gamma\lambda$. Therefore,

$$1 - P(\Omega_1(\lambda)) \leq P(\max_{j \in S_3} \|\tilde{\theta}_{nj} - \theta_{nj}\| \geq \theta_* - \gamma\lambda).$$

We also have

$$1 - P(\Omega_2(\lambda)) \leq P(n^{-1} \max_{j \notin S_3} \|Z'_j H(\varepsilon_n + \delta_n)\| > \lambda).$$

Note that when $m_n = n^{1/(2d+1)}$, we have $m_n n^{-1/2} = m_n^{-(2d-1)/2}$. Therefore, with Lemmas 1 and 2, we have $P(\hat{\theta}_n \neq \tilde{\theta}_n) \rightarrow 0$.

Proof of Theorem 2 By the definition of $\hat{\theta}_n = (\hat{\theta}'_{n1}, \dots, \hat{\theta}'_{np})'$,

$$\frac{1}{2n} \|Q(y - Z\hat{\theta}_n)\|_2^2 + \sum_{j=1}^p \rho(\|\hat{\theta}_{nj}\|; \lambda, \gamma) \leq \frac{1}{2n} \|Q(y - Z\theta_n)\|_2^2 + \sum_{j=1}^p \rho(\|\hat{\theta}_{nj}\|; \lambda, \gamma). \quad (\text{A.8})$$

Let $\eta_n = Q(y - Z\theta_n)$, $\nu_n = QZ(\hat{\theta}_n - \theta_n)$, and write

$$Q(y - Z\hat{\theta}_n) = Q(y - Z\theta_n) - QZ(\hat{\theta}_n - \theta_n) = \eta_n - \nu_n.$$

We have $\|Q(y - Z\hat{\theta}_n)\|_2^2 = \|\nu_n\|_2^2 - 2\eta_n^T \nu_n + \|\eta_n\|_2^2$. We can rewrite as

$$\|\nu_n\|_2^2 - 2\eta_n^T \nu_n \leq 2n \sum_{j=1}^p (\rho(\|\theta_{nj}\|; \lambda, \gamma) - \rho(\|\hat{\theta}_{nj}\|; \lambda, \gamma)). \quad (\text{A.9})$$

Since

$$|\rho(\|\theta_{nj}\|; \lambda, \gamma) - \rho(\|\hat{\theta}_{nj}\|; \lambda, \gamma)| \leq \lambda \|\theta_{nj} - \hat{\theta}_{nj}\|. \quad (\text{A.10})$$

combining (A.9) and (A.10), we get

$$\|\nu_n\|_2^2 - 2\eta_n^T \nu_n \leq 2n\lambda\sqrt{p}\|\hat{\theta}_n - \theta_n\|. \quad (\text{A.11})$$

Let $\eta_n^* = QZ(Z'QZ)^{-1}Z'Q\eta_n$. By the Cauchy-Schwartz inequality,

$$2|\eta_n^T \nu_n| \leq 2\|\eta_n^*\|_2 \cdot \|\nu_n\|_2 \leq 2\|\eta_n^*\|_2^2 + \frac{1}{2}\|\nu_n\|_2^2. \quad (\text{A.12})$$

From (A.11) and (A.12), we have

$$\|\nu_n\|_2^2 \leq 4\|\eta_n^*\|_2^2 + 4n\lambda\sqrt{p} \cdot \|\hat{\theta}_n - \theta_n\|_2.$$

Let c_* be the smallest eigenvalue of $Z^T QZ/n$. By Lemma 1 of Huang, Horowitz, and Wei (2010), $c_* \asymp_p m_n^{-1}$. Since $\|\nu_n\|_2^2 \geq nc_* \|\hat{\theta}_n - \theta_n\|_2^2$ and $2ab \leq a^2 + b^2$,

$$nc_* \|\hat{\theta}_n - \theta_n\|_2^2 \leq 4\|\eta_n^*\|_2^2 + \frac{8\|2n\lambda\sqrt{p}\|_2^2}{2nc_*} + \frac{1}{2}nc_* \|\hat{\theta}_n - \theta_n\|_2^2.$$

It follows that

$$\|\hat{\theta}_n - \theta_n\|_2^2 \leq \frac{8\|\eta_n^*\|_2^2}{nc_*} + \frac{4\lambda^2 p}{c_*^2}. \quad (\text{A.13})$$

Let $\alpha(u_i) = \sum_{j=1}^p \alpha_{0j}(u_i)$. Write

$$\eta = Q(\varepsilon_i + \alpha(u_i) - Z\theta).$$

Since $\|\alpha_{0j}(u) - \alpha_j(u)\| = O(m_n^{-d})$, we have

$$\|\eta_n^*\|_2^2 \leq 2\|\varepsilon_n^*\|_2^2 + O(np m_n^{-2d}). \quad (\text{A.14})$$

where ε_n^* is the projection of $\varepsilon_n = (\varepsilon_1, \dots, \varepsilon_n)'$ to the span of QZ . We have

$$\|\varepsilon_n^*\|_2^2 = \|(Z'QZ)^{-1/2} Z'Q\varepsilon_n\|_2^2 \leq O_p(p m_n). \quad (\text{A.15})$$

Combining (A.13), (A.14) and (A.15), we get

$$\|\hat{\theta}_n - \theta_n\|_2^2 \leq O_p\left(\frac{p m_n}{n c_*}\right) + O\left(\frac{d_2 m_n^{-2d}}{c_*}\right) + \frac{4p\lambda^2}{c_*^2}.$$

Since $c_* \asymp_p m_n^{-1}$ and $c^* \asymp_p m_n^{-1}$, we have

$$\|\hat{\theta}_n - \theta_n\|_2^2 \leq O_p\left(\frac{p m_n^2}{n}\right) + O\left(\frac{1}{m_n^{2d-1}}\right) + O(m_n^2 \lambda^2).$$

Proof of Theorem 3 Let $\tilde{\theta}_n$ be the oracle estimator defined in (3.1), and let

$$\tilde{g}_{nj}(u) = 0, \quad j \in S_1 \cup S_2 \quad \text{and} \quad \tilde{g}_{nj}(u) = \sum_{k=1}^{m_n} \tilde{\theta}_{jk} \psi_{jk}(u), \quad j \in S_3,$$

$$\tilde{\alpha}_{nj}(u) = (1, u) \tilde{\beta}_{nj} + \tilde{g}_{nj}(u), \quad j \in \hat{S}_3.$$

The estimator of the coefficients of the linear components is

$$\tilde{\beta}_{n1} = (X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T (y - X \tilde{\alpha}(u)).$$

Using standard techniques in semiparametric models, such as those described in Huang (1996), we can show that

$$\sqrt{n}(\tilde{\beta}_{n1} - \beta_{01}) \xrightarrow{D} N(0, \Sigma).$$

By Theorem 1, $P(\hat{\beta}_{n1} = \tilde{\beta}_{n1}) \rightarrow 1$, which implies $\sqrt{n}(\hat{\beta}_{n1} - \tilde{\beta}_{n1}) \xrightarrow{P} 0$. Therefore, by Slutsky's Lemma, we also have

$$\sqrt{n}(\tilde{\beta}_{n1} - \beta_{01}) = \sqrt{n}(\tilde{\beta}_{n1} - \beta_{01}) + \sqrt{n}(\hat{\beta}_{n1} - \tilde{\beta}_{n1}) \xrightarrow{D} N(0, \Sigma).$$

This completes the proof of Theorem 3.

References

- Breheny P and Huang J (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Statist.* 5, 232–253.
- Chiang CT, Rice JA and Wu CO (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Amer. Statist. Assoc.* 96, 605–619.
- Cleveland WS, Grosse E and Shyu WM (1991). Local regression models In *Statistical Models in S* (Chambers JM and Hastie TJ, eds), 309–376. Wadsworth & Brooks, Pacific Grove.
- Fan J and Huang T (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11, 1031–1057.
- Fan J and Zhang W (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* 27, 1491–1518.
- Fu WJ (1998). Penalized regressions: the bridge versus the LASSO. *J. Comp. Graph. Statist.* 7, 397–416.
- Hastie TJ and Tibshirani RJ (1993). Varying-coefficient models. *J. Roy. Statist. Soc. B.* 55, 757–796.
- Hoover DR, Rice JA, Wu CO and Yang LP (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85, 1048–1062.
- Huang JZ, Wu CO and Zhou L (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89, 111–128.
- Huang JZ, Wu CO and Zhou L (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Stat Sinica*. 14, 763–788.
- Huang JZ and Shen H (2004). Functional coefficient regression models for nonlinear time series: A polynomial spline approach. *Scand J. Stat.* 31, 515–534.
- Huang J, Breheny P and Ma S, (2012). A selective review of group selection in high dimensional models. *Statist. Sci.* 27, 481–499.
- Huang J, Wei FR and Ma S, (2012). Semiparametric regression pursuit. *Stat. Sinica*, 22, 1403–1426.
- Huang JH, Wu CO and Zhou L (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Stat. Sinica*, 14, 763–788.
- Kauermann G and Tutz G (1999). On model diagnostics using varying-coefficient models. *Biometrika* 86, 119–128.
- Lam C and Fan J (2008). Profile-kernel likelihood inference with diverging number of parameters. *Ann. Statist.* 36, 2232–2260.
- Li R and Liang H (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.* 36, 261–286.
- Meier L, Van de Geer S and Bulmann P (2008). The group lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B*, 70, 53–71.
- Schumaker L (1981). *Spline Functions: basic theory*. Wiley, New York.
- Schwarz G (1978). Estimating the dimension of a model. *Ann. Statist.* 6, 461–464.

- Wang H and Xia Y (2008). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.*, 104, 747–757.
- Wei FR, Huang J and Li H (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Stat. Sinica*, 21, 1515–1540.
- Wu CO, Chiang CT and Hoover DR (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Amer. Statist. Assoc.* 93, 1388–1402.
- Yuan M and Lin Y (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B*, 68, 49–67.
- Zhang CH (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* 38, 894–942.

Table 1:

Simulation results for Examples 1-2. NL, the average number of the nonlinear components being selected; ER, the average model error; IN%, the percentage of occasions in which the correct nonlinear components are included in the selected model; CS%, the percentage of occasions in which exactly correct nonlinear components are selected, averaged over 100 replications. Enclosed in parentheses are the corresponding standard errors.

	n=400				n=800			
	NL	ER	IN%	CS%	NL	ER	IN%	CS%
Example 1, group Lasso	14.44 (1.506)	0.975 (0.076)	100 (0.000)	100 (0.000)	12.5 (1.521)	1.092 (0.061)	100 (0.000)	100 (0.000)
group MCP	7.14 (2.142)	0.955 (0.081)	100 (0.000)	100 (0.000)	4.28 (0.817)	1.069 (0.062)	100 (0.000)	100 (0.000)
Example 2, group Lasso	11.83 (1.341)	1.112 (0.166)	32 (0.469)	32 (0.469)	12.97 (1.396)	1.098 (0.068)	35 (0.479)	35 (0.479)
group MCP	6.71 (1.358)	1.044 (0.165)	77 (0.423)	77 (0.423)	6.07 (1.328)	1.079 (0.067)	95 (0.219)	95 (0.219)

Table 2:

Number of times each component was selected as a nonlinear component by the group Lasso and group MCP methods in the 100 replications, in Example 1.

	$\alpha_1(\cdot)$	$\alpha_2(\cdot)$	$\alpha_3(\cdot)$	$\alpha_4(\cdot)$
n=400				
Example 1, group Lasso	0.112 (0.066)	0.296 (0.062)	1.748 (0.179)	0.933 (0.299)
group MCP	0.109 (0.095)	0.407 (0.141)	1.968 (0.275)	1.712 (0.513)
n=800				
group Lasso	0.074 (0.042)	0.280 (0.037)	1.700 (0.138)	1.005 (0.180)
group MCP	0.068 (0.048)	0.348 (0.075)	1.905 (0.207)	1.621 (0.270)

Table 3:

Number of times each component was selected as a nonlinear component by the group Lasso and group MCP methods in the 100 replications, in Example 2.

	$\alpha_1(\cdot)$	$\alpha_2(\cdot)$	$\alpha_3(\cdot)$	$\alpha_4(\cdot)$	$\alpha_5(\cdot)$	$\alpha_6(\cdot)$
n=400						
Example 2, group Lasso	0.171 (0.091)	0.301 (0.084)	1.676 (0.163)	0.845 (0.272)	18.159 (0.122)	2.492 (0.210)
group MCP	0.086 (0.072)	0.381 (0.172)	1.918 (0.261)	1.705 (0.467)	18.170 (0.217)	1.992 (0.338)
n=800						
group Lasso	0.070 (0.049)	0.267 (0.042)	1.780 (0.135)	1.122 (0.206)	18.022 (0.113)	2.263 (0.169)
group MCP	0.079 (0.057)	0.358 (0.114)	1.960 (0.195)	1.794 (0.339)	18.061 (0.177)	1.841 (0.159)