Statistica Sinica **22** (2012), 1403-1426
doi:http://dx.doi.org/10.5705/ss.2010.298

SEMIPARAMETRIC REGRESSION PURSUIT

Jian Huang, Fengrong Wei and Shuangge Ma

University of Iowa, University of West Georgia and Yale University

Abstract: The semiparametric partially linear model allows flexible modeling of covariate effects on the response variable in regression. It combines the flexibility of nonparametric regression and the parsimony of linear regression. The most important assumption in the existing methods for the estimation in this model is that a priori it is known which covariates have a linear effect and which do not. However, in applied work, this is rarely known in advance. We consider the problem of estimation in the partially linear models without assuming a priori which covariates have linear effects. We propose a semiparametric regression pursuit method for identifying the covariates with a linear effect. Our proposed method is a penalized regression approach using a group minimax concave penalty. Under suitable conditions we show 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. The performance of the proposed method is evaluated using simulation studies that support our theoretical results. A data example is used to illustrated the application of the proposed method.

Key words and phrases: Group selection, minimax concave penalty, model-pursuit consistency, penalized regression, semiparametric models, structure estimation.

1. Introduction

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

$$y_i = \mu + \sum_{j \in S_1} \beta_j x_{ij} + \sum_{j \in S_2} f_j(x_{ij}) + \varepsilon_i, 1 \le i \le n, \tag{1.1}$$

where S_1 and S_2 are mutually exclusive and complementary subsets of $\{1, \ldots, p\}$, $\{\beta_j : j \in S_1\}$ are regression coefficients of the covariates with indices in S_1, and $(f_j : j \in S_2)$ are unknown functions. In this model, the mean response is linearly related to the covariates in S_1, while its relation with the remaining covariates is not specified up to any finite number of parameters. This model combines the flexibility of nonparametric regression and the parsimony of linear regression. When the relation between y_i and $\{x_{ij} : j \in S_1\}$ is of main interest and can be

approximated by a linear function, it offers more interpretability than a purely nonparametric additive model.

There is a large literature on estimation in partially linear models. Examples include the partial spline estimator (Wahba (1984), Engle et al. (1986), and Heckman (1986)), the partial residual estimator (Robinson (1988), Speckman (1985)), and polynomial spline estimator (Chen (1988)). An excellent discussion of partially linear models can be found in the book of Härdle Liang, and Gao (2000), which also contains an extensive list of references on this model. A comprehensive treatment of general semiparametric theory and many related models can be found in Bickel et al. (1993).

The most important assumption in the existing methods for the estimation in partially linear models is that it is known a priori which covariates have a linear form and which do not in the model. This assumption underlies the construction of the estimators and investigation of their theoretical properties in the existing methods. However, in applied work, it is rarely known in advance which covariates have linear effects and which have nonlinear effects.

Recently, Zhang, Cheng, and Liu (2011) proposed a novel method for determining the zero, linear, and nonlinear components in partially linear models. Their method is a two-step regularization method in the smoothing spline ANOVA framework. In the first step, they obtain an initial consistent estimator for the components in a nonparametric additive model, and then use the initial estimator as the weights in their proposed regularized smoothing spline method in a way similar to the adaptive Lasso (Zou (2006)). They obtained the rate of convergence of their proposed estimator. They also showed that their method is selection consistent in the special case of tensor product design. However, they did not prove any selection consistency results for general partially linear models. Also, in their two-step approach, a total of four penalty parameters need to be selected, and this may be difficult to implement in practice.

We consider the problem of estimation in partially linear models without assuming a priori which covariates have linear effects and which have nonlinear effects, and propose a semiparametric regression pursuit method for identifying them. We embed partially linear models into a nonparametric additive model. By approximating the nonparametric components using spline series expansions, we transform the problem of model specification into a group variable selection problem. We then determine the linear and nonlinear components with a penalized approach, using a minimax concave penalty (MCP, Zhang (2010)) imposed on the norm of the coefficients in the spline expansion. We refer to this penalized approach as the group MCP method. We show that, under suitable conditions, 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. We allow the possibility that the underlying true model is not partially

linear. Then the proposed approach has the same asymptotic property as the nonparametric estimator in the nonparametric additive model. We also show that the estimated coefficients of linear effects are asymptotically normal, with the same distribution as the estimator assuming the true model were known in advance.

Some of the techniques used in this paper are similar to those in Huang, Horowitz, and Wei (2010), in which the problem of variable selection in nonparametric additive models is considered. In particular, after transforming the present problem of model pursuit into a group selection problem based on spline approximation, some of the techniques in obtaining rate of convergence for the group Lasso estimator in the context of nonparametric additive models in Huang, Horowitz, and Wei (2010) can be applied, with some modifications, see the proof of Theorem 2 in the Appendix. However, the problem of model pursuit considered here is very different from that in Huang, Horowitz, and Wei (2010). Also, we use the group MCP rather than the group Lasso, which requires different treatment at the technical level.

This article is organized as follows. In Section 2 we describe our proposed semiparametric regression pursuit (SRP) method. We transform the problem of identifying linear and nonlinear components into an group selection problem using the group MCP. In Section 3 we derived a group coordinate descent algorithm to implement the proposed method. In Section 4 we state the theoretical results concerning the selection and estimation properties of the proposed method. Section 5 includes simulation studies and an illustration of the proposed method on a data example. Proofs of the results stated in Section 3 are given in the Appendix.

2. Semiparametric Regression Pursuit via Group Minimax Concave Penalization

2.1. Method

The semiparametric partially linear model (1.1) can be embedded into the nonparametric additive model (Hastie and Tibshirani (1990)),

$$y_i = \mu + f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \varepsilon_i. \tag{2.1}$$

Suppose that x_{ij} takes values in $[a, b]$, where $a < b$ are finite constants. To ensure unique identification of the f_j's, we take $\mathrm{E}f_j(x_{ij}) = 0, 1 \le j \le p$. If some of the f_j's are linear, then (2.1) becomes the partially linear additive model (1.1). The problem is that of determining which f_j's have a linear form and which do not. For this purpose, we write f_j as

$$f_j(x) = \beta_{0j} + \beta_j x + g_j(x).$$

Consider a truncated series expansion for approximating g_j,

$$g_{nj}(x) = \sum_{k=1}^{m_n} \theta_{jk}\phi_k(x), \tag{2.2}$$

where $\phi_1, \ldots, \phi_{m_n}$ are basis functions and $m_n \to \infty$ at a certain rate as $n \to \infty$. If $\theta_{jk} = 0, 1 \le k \le m_n$, then f_j is linear. Therefore, with this formulation, the problem is to determine which groups of $\{\theta_{jk}, 1 \le k \le m_n\}$ are zero.

Let $\beta = (\beta_1, \ldots, \beta_p)'$ and $\theta_n = (\theta_{1n}', \ldots, \theta_{pn}')'$, where $\theta_{jn} = (\theta_{j1}, \ldots, \theta_{jm_n})'$. Define the penalized least squares criterion

$$\begin{aligned} L(\mu, \beta, \theta_n; \lambda, \gamma) = {} & \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j - \sum_{j=1}^{p}\sum_{k=1}^{m_n}\theta_{jk}\phi_k(x_{ij})\Big)^2 \\ & + \sum_{j=1}^{p} \rho_\gamma(\|\theta_{jn}\|_{A_j}; \sqrt{m_n}\lambda), \end{aligned} \tag{2.3}$$

where ρ is a penalty function depending on the penalty parameter $\lambda \ge 0$ and a regularization parameter γ. Here without causing confusing, we still use μ to denote the intercept. The norm $\|\theta_{jn}\|_{A_j} = (\theta_{nj}' A_j \theta_{nj})^{1/2}$ for a given positive definite matrix A_j. In theory, any positive definite matrix can be used as A_j, since $\|\theta_{jn}\|_{A_j} = 0$ if and only if $\theta_{jn} = 0$. 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 will specify A_j in (2.8) below.

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

$$\rho_\gamma(t; \lambda) = \lambda \int_0^t \Big(1 - \frac{x}{\gamma\lambda}\Big)_+ dx, \ t \ge 0, \tag{2.4}$$

where γ is a parameter that controls the concavity of ρ and λ is the penalty parameter. Here $x_+ = x1_{\{x \ge 0\}}$. We require $\lambda \ge 0$ 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 MCP can be easily understood by considering its derivative

$$\dot{\rho}_\gamma(t; \lambda) = \lambda\Big(1 - \frac{t}{\gamma\lambda}\Big)_+, \ t \ge 0. \tag{2.5}$$

It begins by applying the same rate of penalization as the lasso, but continuously relaxes that penalization until, when $t > \gamma\lambda$, the rate of penalization drops to 0. It provides a continuum of penalties with the ℓ_1 penalty at $\gamma = \infty$ and the hard-thresholding penalty as $\gamma \to 1+$. In particular, it includes the Lasso penalty

as a special case at $\gamma = \infty$. Detailed discussions on the MCP can be found in Zhang (2010).

The penalty at (2.3) is a composite of the penalty function $\rho_\gamma(\cdot;\lambda)$ and a weighted ℓ_2-norm of θ_j. The $\rho_\gamma(\cdot;\lambda)$ 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 (2.3), it is natural to call it the group minimax concave penalty, or group MCP.

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

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

subject to the constraints

$$\sum_{i=1}^{n}\sum_{k=1}^{m_n} \theta_{jk}\phi_k(x_{ij}) = 0, 1 \le j \le p. \tag{2.6}$$

These centering constraints are sample analogs of the identifying restriction $Ef_j(x_{ij}) = 0, 1 \le i \le n, 1 \le j \le p$.

We convert (2.6) to an unconstrained optimization problem by centering the response and the covariate functions. Specifically, we center the responses and covariates, and standardize the covariates by setting

$$\sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij} = 0 \ \text{ and } \ \sum_{i=1}^{n} x_{ij}^2 = n.$$

We also center the basis functions and take

$$\bar{\phi}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \phi_k(x_{ij}), \quad \psi_{jk}(x) = \phi_k(x) - \bar{\phi}_{jk}. \tag{2.7}$$

Let $z_{ij} = \big(\psi_{j1}(x_{ij}), \ldots, \psi_{jm_n}(x_{ij})\big)'$, which consists of the centered basis functions at the ith observation of the jth covariate. Let $Z = (Z_1, \ldots, Z_p)$, where $Z_j = (z_{1j}, \ldots, z_{nj})'$ is the $n \times m_n$ 'design' matrix corresponding to the jth expansion. Let $y = (y_1, \ldots, y_n)'$, $x_j = (x_{1j}, \ldots, x_{nj})'$ and $X = (x_1, \ldots, x_p)$. We can write

$$(\hat{\beta}_n, \hat{\theta}_n) = \arg\min_{\beta,\theta_n}\{L(\beta, \theta_n; \lambda, \gamma) = \frac{1}{2n}\|y - X\beta - Z\theta_n\|^2 + \sum_{j=1}^{p} \rho_\gamma(\|\theta_{nj}\|_{A_j}; \sqrt{m_n}\lambda)\}.$$

Here we dropped μ from the arguments of L, since the intercept is zero due to centering. With the centering, the constrained optimization problem becomes an unconstrained one.

2.2. Penalized profile least squares

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 $\beta = (X'X)^{-1}X'(y - Z\theta_n)$. 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

$$L(\theta_n; \lambda, \gamma) = \frac{1}{2n}\|Q(y - Z\theta_n)\|^2 + \sum_{j=1}^{p} \rho_\gamma(\|\theta_{nj}\|_{A_j}; \sqrt{m_n}\lambda). \tag{2.8}$$

As noted above, any positive definite matrix can be used for A_j. Here we use $A_j = Z_j'QZ_j/n$. The rationale for this choice is based on the following considerations. First, in the profile objective function (2.8), the covariate matrix for group j is QZ_j. The Gram matrix associated with it is $Z_j'Q'QZ_j/n = A_j$, since Q is an orthonormal matrix. Although the original covariates x_{ij}'s are standardized, the covariate matrices for the groups are not necessarily so. 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 described 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 $\hat{\theta}_n = \arg\min_{\theta_n} L(\theta_n; \lambda, \gamma)$. We compute $\hat{\theta}_n$ using the group coordinate descent algorithm described in Section 3.

The set of indices of the covariates that are estimated to have the linear form in the regression model (1.1) is $\hat{S}_1 \equiv \{j : \|\hat{\theta}_{nj}\| = 0\}$. Thus,

$$\hat{g}_{nj}(x) = 0, j \in \hat{S}_1 \quad \text{and } \hat{g}_{nj}(x) = \sum_{k=1}^{m_n} \hat{\theta}_{jk}\psi_{jk}(x), j \notin \hat{S}_1.$$

Let $\widehat{X}_{(1)} = (x_j, j \in \hat{S}_1)$, $\widehat{Z}_{(2)} = (Z_j : j \notin \hat{S}_1)$ and $\hat{\theta}_{n(2)} = (\hat{\theta}_{nj}', j \notin \hat{S}_1)'$. We have $\hat{\beta}_n = (X'X)^{-1}X'(y - \widehat{Z}_{(2)}\hat{\theta}_{n(2)})$. The estimator of the coefficients of the linear components is $\hat{\beta}_{n1} = (\hat{\beta}_j : j \in \hat{S}_1)'$. Let

$$\hat{f}_{nj}(x) = \hat{\beta}_j x + \hat{g}_{nj}(x), \ j \notin \hat{S}_1.$$

Write $\hat{f}_{nj}(x_j) = (\hat{f}_{nj}(x_{1j}), \ldots, \hat{f}_{nj}(x_{nj}))'$. Then the estimator of the coefficient vector of the linear components can also be written as

$$\hat{\beta}_{n1} = (\widehat{X}_{(1)}'\widehat{X}_{(1)})^{-1}\widehat{X}_{(1)}'(y - \sum_{j \notin \hat{S}_1} \hat{f}_{nj}(x_j)).$$

2.3. Spline approximation

We use polynomial splines to approximate the nonparametric components $g_j, 1 \le j \le p$. Let $a = t_0 < t_1 < \cdots < t_K < t_{K+1} = b$ be a partition of $[a, b]$ into K subintervals $I_{Kk} = [t_k, t_{k+1}), k = 0, \ldots, K-1$ and $I_{KK} = [t_K, t_{K+1}]$, where $K \equiv K_n = O(n^v)$, $0 < v < 0.5$, is a positive integer such that $\max_{1\le k\le K+1} |t_k - t_{k-1}| = O(n^{-v})$. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \ge 1$ consisting of functions s for which the restriction of s to I_{Kk} is a polynomial of degree l for $1 \le k \le K$, and for $l \ge 2$ and $0 \le l' \le l-2$, s is l' times continuously differentiable on $[a, b]$ (Schumaker (1981)). There exists normalized B-spline basis functions $\{\phi_k, 1 \le k \le m_n\}$ for $\mathcal{S}_n$, where $m_n \equiv K_n + l$ (Schumaker (1981)). We can use these basis functions in the approximation (2.2).

3. Computation

We derive a group coordinate descent algorithm for computing $\hat{\theta}_n$. This algorithm is a natural extension of the standard coordinate descent algorithm (Fu (1998), Friedman et al. (2007), and Wu and Lange (2007)) used in optimization problems with convex penalties, such as the Lasso. It has also been used in calculating the penalized estimates based on concave penalty functions (Breheny and Huang (2011), Mazumder, Friedman, and Hastie (2009)).

The group coordinate descent algorithm optimizes a target function with respect to a single group at a time, iteratively cycling through all groups until convergence is reached. It is particularly suitable for computing $\hat{\theta}_n$, since it has a simple closed form expression for a single-group model, see (3.1) below.

We write $A_j = R_j' R_j$ for an $m_n \times m_n$ upper triangular matrix R_j via the Cholesky decomposition. Let $b_j = R_j \theta_j$, $\tilde{y} = Qy$ and $\tilde{Z}_j = QZ_j R_j^{-1}$. Simple algebra shows that

$$L(b; \lambda, \gamma) = \frac{1}{2n} \|\tilde{y} - \sum_{j=1}^{p} \tilde{Z}_j b_j\|^2 + \sum_{j=1}^{p} \rho_\gamma(\|b_j\|; \sqrt{m_n}\lambda).$$

Note that $n^{-1}\tilde{Z}_j'\tilde{Z}_j = R_j^{-1\prime}(n^{-1}Z_j'QZ_j)R_j^{-1} = I_{m_n}$. Let $\tilde{y}_j = \tilde{y} - \sum_{k\ne j}^{p} \tilde{Z}_k b_k$ and

$$L_j(b_j; \lambda, \gamma) = \frac{1}{2n} \|\tilde{y}_j - \tilde{Z}_j b_j\|^2 + \rho_\gamma(\|b_j\|; \sqrt{m_n}\lambda).$$

Let $\eta_j = \tilde{Z}_j(\tilde{Z}_j'\tilde{Z}_j)^{-1}\tilde{y}_j = n^{-1}\tilde{Z}_j'\tilde{y}$. For $\gamma > 1$, it can be verified that the value that minimizes L_j with respect to b_j is

$$\tilde{b}_{j,GM}(\lambda, \gamma) = M(\eta_j; \lambda, \gamma) \equiv \begin{cases} 0, & \text{if } \|\eta_j\| \le \sqrt{m_n}\lambda, \\ \frac{\gamma}{\gamma - 1}(1 - \frac{\sqrt{m_n}\lambda}{\|\eta_j\|})\eta_j, & \text{if } \sqrt{m_n}\lambda < \|\eta_j\| \le \gamma\sqrt{m_n}\lambda, \\ \eta_j, & \text{if } \|\eta_j\| > \gamma\sqrt{m_n}\lambda. \end{cases} \tag{3.1}$$

In particular, when $\gamma = \infty$, we have

$$\tilde{b}_{j,GL} = \Big(1 - \frac{\sqrt{m_n}\lambda}{\|\eta_j\|}\Big)_+ \eta_j,$$

which is the group Lasso estimate for a single-group model (Yuan and Lin (2006)).

The group coordinate descent algorithm can now be implemented as follows. Suppose the current values for the group coefficients $\tilde{b}_k^{(s)}, k \neq j$ are given. We want to minimize L with respect to b_j. Let

$$L_j(b_j; \lambda, \gamma) = \frac{1}{2n}\|\tilde{y} - \sum_{k \neq j} \tilde{Z}_k \tilde{b}_k^{(s)} - \tilde{Z}_j b_j\|^2 + \rho_\gamma(\|b_j\|; \sqrt{m_n}\lambda),$$

and write $\tilde{y}_j = \sum_{k \neq j} \tilde{Z}_k \tilde{b}_k^{(s)}$ and $\tilde{\eta}_j = n^{-1}\tilde{Z}_j'(\tilde{y} - \tilde{y}_j)$. Let $\tilde{b}_j$ be the minimizer of $L_j(b_j; \sqrt{m_n}\lambda, \gamma)$. When $\gamma > 1$, we have $\tilde{b}_j = M(\tilde{\eta}_j; \sqrt{m_n}\lambda, \gamma)$, where M is defined in (3.1).

For any given (λ, γ), we use (3.1) to cycle through one component at a time. Let $\tilde{\beta}^{(0)} = (\tilde{\beta}_1^{(0)\prime}, \ldots, \tilde{\beta}_p^{(0)\prime})'$ be the initial value. The proposed coordinate descent algorithm is as follows.

Initialize vector of residuals $r = y - \tilde{y}$, where $\tilde{y} = \sum_{j=1}^p \tilde{Z}_j b_j^{(0)}$. For $s = 0, 1, \ldots$, carry out the following calculation until convergence. For $j = 1, \ldots, p$, repeat the following steps.

(1) Calculate $\tilde{\eta}_j = n^{-1}\tilde{Z}_j' r + \tilde{b}_j^{(s)}$.

(2) Update $\tilde{b}_j^{(s+1)} = M(\tilde{\eta}_j; \lambda, \gamma)$.

(3) Update $r \leftarrow r - \tilde{Z}_j(\tilde{b}_j^{(s+1)} - \tilde{b}_j^{(s)})$ and $j \leftarrow j + 1$.

The last step ensures that r holds the current values of the residuals. Although the objective function is not necessarily convex, it is convex with respect to a single group when the coefficients of all the other groups are fixed. Thus, Theorem 5.1 of Tseng (2001) implies that the group coordinate descent algorithm described above converges.

4. Theoretical Properties

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

Denote the underlying regression components by f_{0j} and write

$$f_{0j}(x) = \beta_{0j} x + g_{0j}(x).$$

Suppose the series expansion for approximating g_{0j} is

$$g_{0j}(x) = \sum_{j=1}^{m_n} \theta_{0jk}\phi_k(x).$$

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

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

$$\tilde{\theta}_n = \arg\min_{\theta_n}\{\frac{1}{2n}\|Q(y - Z\theta_n)\|^2 : \theta_{nj} = 0, j \in S_1\}. \tag{4.1}$$

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 2.2, write the oracle estimators as

$$\tilde{g}_{nj}(x) = 0, j \in S_1 \text{ and } \tilde{g}_{nj}(x) = \sum_{k=1}^{m_n} \tilde{\theta}_{jk}\psi_{jk}(x), j \notin S_1.$$

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

$$\tilde{f}_{nj}(x) = \tilde{\beta}_j x + \tilde{g}_{nj}(x), \ j \in S_2.$$

Write $\tilde{f}_{nj}(x_j) = (\tilde{f}_{nj}(x_{1j}), \ldots, \tilde{f}_{nj}(x_{nj}))'$. The oracle estimator of the coefficients of the linear components is

$$\tilde{\beta}_{n1} = (X'_{(1)}X_{(1)})^{-1}X'_{(1)}(y - \sum_{j\in S_2} \tilde{f}_{nj}(x_j)).$$

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

$$\tilde{\theta}_{n(2)} = (Z'_{(2)}QZ_{(2)})^{-1}Z'_{(2)}Qy. \tag{4.2}$$

Let $\theta_* = \min_{j\in S_1} \|\theta_{0nj}\|$, 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 kth derivative $g^{(k)}$ exists and satisfies a Lipschitz condition of order α:

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

We make the following assumptions.

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

(A2) $\mathrm{E}g_j(x_j) = 0$ and $g_j \in \mathcal{G}, j = q+1, \ldots, p$.

(A3) The covariate vector X has a continuous density and there exist constants C_1 and C_2 such that the density function η_j of x_j satisfies $0 < C_1 \le \eta_j(x) \le C_2 < \infty$ on $[a,\ b]$ for every $1 \le j \le p$.

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*

$$\frac{1}{m_n^{(2d-1)/2}(\theta_* - \gamma\lambda)} + \frac{1}{\lambda\sqrt{n}} \to 0. \tag{4.3}$$

Then under (A1)−(A3), $\mathrm{P}(\hat{\theta}_n \ne \tilde{\theta}_n) \to 0$. *Consequently,* $\mathrm{P}(\hat{S}_1 = S_1) \to 1$, *and*

$$\mathrm{P}(\hat{\beta}_{n1} = \tilde{\beta}_{n1}) \to 1, \quad \textit{and} \quad \mathrm{P}(\|\hat{f}_{nj} - \tilde{f}_{nj}\|_2 = 0, j \in S_2) \to 1.$$

Therefore, under the conditions of Theorem 1, the proposed estimator can correctly distinguish 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 linear and nonlinear components were known, except on an event with probability tending to zero.

We note that, except the assumption on the tail probabilities in (A1), (A1)−(A3) are standard conditions for nonparametric additive models. They are needed to estimate the additive components at the optimal ℓ_2 rate of convergence in standard nonparametric additive model setting. The main extra condition here is (4.3), which requires both $\lambda = o(n^{-1/2})$ and $\theta_* > \gamma\lambda + a_n m_n^{-(2d-1)/2}$ for some $a_n \to \infty$. 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. *Suppose* (A1)−(A3) *hold. Under* (2.1), *we have*

$$\sum_{j=1}^{p} \|\hat{f}_{nj} - f_{0j}\|_2^2 \le O_p\Big(\frac{m_n}{n}\Big) + O\Big(\frac{1}{m_n^{2d}}\Big) + O(m_n\lambda^2).$$

This theorem gives rate of convergence of the proposed estimator under the nonparametric additive model (2.1) that contains the partially linear models as

special cases. In particular, if we assume that each component in (2.1) 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 $\sum_{j=1}^{p} \|\hat{f}_{nj} - f_{0j}\|_2^2 = O_p(n^{-4/5})$, which is the optimal rate of convergence in nonparametric regression.

We now consider the asymptotic distribution of $\hat{\beta}_{n1}$. Let

$$H_j = \{h_j = (h_{jk} : k \in S_1)' : \mathrm{E}h_{jk}^2(x_j) < \infty,\ Eh_{jk}(x_j) = 0\}, \quad j \in S_2.$$

Each element of H_j is a $|S_1|$-vector of square integrable functions with mean zero. Let the sumspace be

$$H = \{h = \sum_{j \in S_2} h_j : h_j \in H_j\}.$$

The projection of the centered covariate vector $x_{(1)} - E(x_{(1)}) \in R^q$ onto the sumspace H is the $(h_1^*, \ldots, h_r^*)'$, with $\mathrm{E}h_j^*(x_j) = 0, j \leq \hat{S}_2$, that minimizes

$$W(h) \equiv \mathrm{E}\|x_{(1)} - \mathrm{E}(x_{(1)}) - \sum_{j \in S_2} h_j(x_j)\|^2. \tag{4.4}$$

For $x_{(2)} = (x_j : j \in S_2)$, write

$$h^*(x_{(2)}) = \sum_{j \in S_2} h_j^*(x_j). \tag{4.5}$$

Under (A3), by Lemma 1 of Stone (1985) and Proposition 2 in Appendix 4 of Bickel et al. (1993), the sumspace H is closed. Thus the orthogonal projection h^* onto H is well defined and unique. Furthermore, each individual component h_j^* is also well-defined and unique. In addition to (A1)-(A3), we also need the following condition.

(A4) Let $w \geq 1$ be a positive integer. The wth partial derivatives of the joint density of $x_{(2)} = (x_j, j \in S_2)$ are bounded by a constant and the qth derivative of each component of $\xi(v) = E(x_{(1)}|x_j = v), j \in S_2$, is bounded by a constant.

Let $A = E[x_{(1)} - E(x_{(1)} - h^*(x_{(2)})]^{\otimes 2}$, where h^* is defined in (4.5). Here $x^{\otimes 2} = xx'$ for any column vector $x \in R^d$.

Theorem 3. *If the conditions of Theorem* 1 *and* (A4) *are satisfied, and if A is nonsingular, then*

$$n^{1/2}(\hat{\beta}_{n1} - \beta_{(1)}) \to_d N(0, \Sigma),$$

where $\beta_{(1)} = (\beta_j : j \in S_1)'$ and $\Sigma = \sigma^2 A^{-1}$.

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

5. Numerical Studies

5.1. Simulation studies

We used simulation to evaluate the finite sample performance of the proposed method. Two examples were considered in the simulation. In each of the simulated models, two sample sizes ($n = 100$, 200) were considered and a total of 100 replications were conducted. Consider the following functions defined on $[0, 1]$:

$$f_1(x) = x,\ f_2(x) = \frac{\sin(2\pi x)}{2 - \sin(2\pi x)},$$
$$f_3(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x),$$
$$f_4(x) = (3x-1)^2,\ f_5(x) = \frac{\cos(2\pi x)}{2 - \cos(2\pi x)},$$
$$f_6(x) = 0.1\cos(2\pi x) + 0.2\sin(2\pi x) + 0.3\cos^2(2\pi x) + 0.4\sin^3(2\pi x) + 0.5\cos^3(2\pi x).$$

In the implementation, we used cubic B-spline with seven basis functions to approximate each function.

Example 1. Let $p = 6$ and consider the model

$$y = 3f_1(x_1) + 4f_1(x_2) - 2f_1(x_3) + 8f_2(x_4) + 6f_3(x_5) + 5f_4(x_6) + \varepsilon.$$

Here the first three variables have linear effect and the last three have nonlinear effect. The p covariates were simulated as follows. First we simulated $w_1, \cdots, w_p$ and u independently from $U[0, 1]$; then $x_{ik} = (w_k + u)/2$ for $k = 1, \cdots, p$. The correlation among predictors was $Corr(x_{ij}, x_{ik}) = 0.5$. The error term ε was chosen from $N(0, 1.57^2)$ to give a signal to noise ratio of 3.

Example 2. Let $p = 10$ and consider the model

$$\begin{aligned} y = {} & 3f_1(x_1) + 4f_1(x_2) - f_1(x_3) - f_1(x_4) + 2f_1(x_5) \\ & + 5f_2(x_6) + 4f_3(x_7) + 5f_4(x_8) + 5f_5(x_9) + 4f_6(x_{10}) + \varepsilon. \end{aligned}$$

Here the first five components are linear and the remaining five are nonlinear. The covariates were simulated as in Example 1. The error term $\varepsilon \sim N(0, 1.80^2)$, which gives a signal to noise ratio of 3.

The group coordinate descent algorithm described in Section 3 was used repeatedly to compute $\hat{\theta}_n$ over a grid of (λ, γ) values in a rectangle $[\lambda_{\max}, \lambda_{\min}] \times [\gamma_{\max}, \gamma_{\min}]$. Here $\lambda_{\max} = \max_{1 \le j \le p} \|n^{-1}\tilde{Z}_j'\tilde{y}\|$, which is the smallest value of λ that forces all the solutions to be zero, and we took $\lambda_{\min} = 0.0001\lambda_{\max}$. We used a set of 100 equally spaced grid points on the logarithmic scale in $[\lambda_{\max}, \lambda_{\min}]$. For

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.

	$n = 100$				$n = 200$			
	NL	ER	IN%	CS%	NL	ER	IN%	CS%
Example 1, Group Lasso	3.46	2.66	100	67	3.10	2.71	100	92
	(0.76)	(0.66)	(0.00)	(0.47)	(0.39)	(0.39)	(0.00)	(0.27)
Group MCP	3.18	2.28	100	82	3.01	2.43	100	99
	(0.39)	(0.47)	(0.00)	(0.39)	(0.10)	(0.30)	(0.00)	(0.10)
Example 2, Group Lasso	4.37	6.26	51	17	5.41	3.55	98	62
	(2.90)	(4.84)	(0.50)	(0.38)	(0.71)	(0.59)	(0.14)	(0.49)
Group MCP	5.25	2.98	76	43	5.22	3.09	98	78
	(1.37)	(1.22)	(0.43)	(0.50)	(0.54)	(0.38)	(0.14)	(0.42)

the γ parameter in the group MCP, we considered a grid of equally spaced points in the interval $[\gamma_{\max}, \gamma_{\min}] = [8.0, 1.1]$ with grid size 0.1. We note that Zhang (2010) suggested using $\gamma = 2.7$ for standardized covariates in linear regression. In our studies, we found that the value of γ has considerable impact on the results. Thus, instead of using a fixed γ value, we considered a range of γ values.

For the group Lasso, which can be considered a special case of the group MCP with $\gamma = \infty$, the algorithm started at $\lambda_{\max}$ where $\hat{\theta}_n$ equals 0 and proceeded along the grid values of λ, using the previous solution as the initial value at each grid point. For the group MCP, for each value of λ in the λ-grid and the corresponding initial value from the group Lasso, the algorithm proceeded along the grids of γ in $[8.0, 1.1]$, that is, for each λ grid value, we started the algorithm at $\gamma = 8$ using the group Lasso solution as the initial value. This approach follows that of Mazumder, Friedman, and Hastie (2009). We then applied BIC (Schwarz (1978)) to select (λ, γ). Here BIC is defined as

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

where $RSS_{\lambda,\gamma}$ is the residual sum of squares and $df_{\lambda,\gamma}$ is the number of the nonzero selected groups for a given (λ, γ). Recall that m_n is the number of spline basis functions given in (2.2). The optimal value of (λ, γ) was chosen to be the one that minimizes the BIC.

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%),

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 Examples 1−2.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
					$n = 100$					
Example 1, Group Lasso	21	13	12	100	100	100				
Group MCP	9	4	5	100	100	100				
					$n = 200$					
Group Lasso	3	4	3	100	100	100				
Group MCP	1	0	0	100	100	100				
					$n = 100$					
Example 2, Group Lasso	19	21	14	17	18	54	73	95	69	57
Group MCP	16	13	9	9	11	89	99	100	97	82
					$n = 200$					
Group Lasso	9	8	7	9	11	99	100	100	100	98
Group MCP	5	6	6	5	2	99	100	100	100	99

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. Table 2 includes the number of times each component estimated as a nonlinear function. Table 3 shows the average mean square error for each function. Enclosed in parentheses are the corresponding standard errors.

Several observations can be made from Tables 1 and 2. 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 $n = 100$, the percentage of correct selection (CS%) was 82% with the group MCP and 67% with the group Lasso. Also, when the sample size was 200, the percentage of inclusion of all the nonlinear components (IN%) and the selection of the correct model (CS%) by both methods were increased. Table 2 shows that the group MCP was more accurate in distinguishing the linear functions from the nonlinear functions than the group Lasso. When $n = 200$, the group MCP correctly distinguished the linear from nonlinear components 99% of the times in Example 1 and 78% of the times in Example 2. In Table 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.

Table 3. The average mean square error for each component selected by the group Lasso and group MCP methods based on 100 replications, in Examples 1−2.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
	$n = 100$									
Example 1, Group Lasso	0.64	0.66	0.67	7.52	12.23	25.50				
	(0.93)	(0.79)	(1.05)	(1.48)	(6.68)	(10.02)				
Group MCP	0.54	0.55	0.49	7.51	11.39	25.34				
	(0.83)	(0.70)	(0.65)	(1.45)	(6.72)	(9.77)				
Oracle	0.11	0.11	0.12	2.22	0.76	10.05				
	(0.25)	(0.17)	(0.23)	(1.07)	(0.46)	(2.39)				
	$n = 200$									
Group Lasso	0.21	0.19	0.20	7.29	12.08	27.24				
	(0.28)	(0.27)	(0.26)	(1.05)	(4.47)	(7.04)				
Group MCP	0.20	0.16	0.19	7.25	11.35	27.08				
	(0.28)	(0.21)	(0.26)	(1.03)	(4.77)	(7.12)				
Oracle	0.09	0.08	0.09	1.88	0.50	9.93				
	(0.07)	(0.06)	(0.07)	(0.65)	(0.18)	(1.72)				
Example 2, Group Lasso	1.22	1.55	1.58	1.40	1.87	3.66	10.24	23.80	3.03	10.09
	(1.45)	(2.63)	(2.08)	(2.06)	(2.95)	(1.43)	(7.17)	(12.7)	(2.76)	(5.80)
Group MCP	0.87	1.05	0.90	0.89	1.03	3.55	9.27	22.30	1.96	9.85
	(1.02)	(1.91)	(1.16)	(1.51)	(1.33)	(1.24)	(6.88)	(10.6)	(1.98)	(5.08)
Oracle	0.52	0.17	0.27	0.31	0.44	2.57	1.09	13.31	1.28	1.85
	(1.00)	(0.60)	(0.36)	(0.63)	(0.79)	(0.90)	(1.54)	(13.9)	(1.80)	(10.45)
	$n = 200$									
Group Lasso	0.34	0.36	0.30	0.38	0.39	3.34	8.55	20.09	0.95	9.26
	(0.45)	(0.40)	(0.41)	(0.61)	(0.56)	(0.71)	(3.19)	(6.61)	(0.81)	(3.86)
Group MCP	0.30	0.32	0.28	0.31	0.34	3.32	8.52	19.91	0.87	9.19
	(0.40)	(0.39)	(0.39)	(0.55)	(0.52)	(0.70)	(3.24)	(6.50)	(0.81)	(3.66)
Oracle	0.23	0.16	0.05	0.16	0.16	0.88	0.36	9.83	0.50	0.33
	(0.20)	(0.23)	(0.02)	(0.33)	(0.41)	(0.30)	(0.14)	(1.68)	(0.17)	(0.14)

5.2. Diabetes data example

This data set is from a study reported in Willems et al. (1997). The data consist of 19 variables on 403 subjects from 1046 African Americans who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia. Diabetes Mellitus Type II (adult onset diabetes) is associated with obesity. The 403 subjects were the ones screened for diabetes. Glycosolated hemoglobin > 7.0 is usually taken as a positive diagnosis of this disease.

We considered Glycosolated hemoglobin as the response variable and the other 15 variables as the covariates. These 15 variables were cholesterol (chol), stabilized glucose (stab.glu), high density lipoprotein (hdl), cholesterol/hdl ratio (ratio), location, age, gender, height, height, weight, frame, first systolic blood

Table 4. Diabetes data: Number of each component being selected by the group Lasso and group MCP methods as nonlinear components. The top panel of Table lists the 12 continuous variables selected by the group MCP and the group Lasso as linear or nonlinear variables, indicated by 0 or 1 (0, linear; 1, nonlinear). The bottom panel shows the number of times a variable had a nonlinear effect in the 100 partitions.

	chol	stab.glu	hdl	ratio	age	height	weight	bp.1s	bp.1d	waist	hip	time.ppn
	whole data set											
group Lasso	0	1	0	0	0	1	0	0	0	0	0	0
group MCP	1	1	0	1	1	1	0	0	0	0	0	1
	training and testing sets											
group Lasso	29	66	7	1	0	72	0	0	0	0	0	0
group MCP	89	100	30	99	65	100	9	2	0	0	4	89

Table 5. Diabetes data: The top panel shows the number of selected nonlinear components (NL) and the residual sum of squares (RSS) based on the whole data. The bottom panel shows the NL, the RSS, and the prediction error (PE), averaged over 100 replications. Enclosed in parentheses are the corresponding standard errors.

	NL	RSS	PE
	whole data		
group Lasso	2.00	3.06	
group MCP	6.00	2.53	
	training and testing sets		
group Lasso	1.75	3.01	3.44
	(0.76)	(0.19)	(1.02)
group MCP	5.87	2.53	3.27
	(0.87)	(0.16)	(0.89)

pressure (bp.1s), first diastolic blood pressure (bp.1d), waist, hip, postprandial time when labs were drawn (time.ppn). Among these 15 variables, three are categorial variables (location, gender, frame) and 12 are continuous variables. We are interested in finding which continuous covariates have nonlinear effects on the response variable. In our study, we only considered the subjects without missing values. Thus the number of subjects was $n = 366$.

The results are summarized in Tables 4 and 5. The top panel of Table 4 lists the 12 continuous variables selected by the group MCP and the group Lasso as linear or nonlinear variables, indicated by 0/1 (1, nonlinear; 0, linear). The top panel of Table 5 shows the number of variables selected as nonlinear variables and the residual sum of squares by both the group MCP and the group Lasso methods.

To evaluate the prediction performance of the methods, we randomly selected a training set with 300 subjects from the data to do the estimation and selection

and used the remaining 66 subjects at the test set for prediction. We repeated this process 100 times and the results are summarized in the bottom panel of Tables 4 and 5. The bottom panel of Table 4 shows the number of times a variable had a nonlinear effect. The bottom panel of Table 5 shows the number of variables being selected (NL) as nonlinear components, the residual sum of squares (RSS) and the prediction error (PE), averaged over 100 replications with standard error in the parentheses. Table 5 shows that the proposed method with the group MCP performed better than the group Lasso in terms of the residual sum of squares and the prediction error.

6. Concluding Remarks

In this paper, we proposed a semiparametric regression pursuit method for distinguishing linear from nonlinear components in semiparametric partially linear models. This approach determines the parametric and nonparametric components in a semiparametric model adaptively based on the data. Our proposed method is fundamentally different from the standard semiparametric inference approach where the parametric and nonparametric components in a model are pre-specified. We showed that our method has the asymptotic oracle properties, meaning that it is the same as the standard semiparametric 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.

We have only considered the proposed semiparametric regression pursuit method in the partially linear model with fixed p. In many applications, such as genomic data analysis, it is possible to have $p > n$. In this case, our proposed method is not directly applicable but, assuming the model is sparse in the sense the number of important covariates is much smaller than n, we can first reduce the model dimension and then apply the proposed method. For example, we can first use the adaptive group Lasso method to select the important variables in the nonparametric additive model (Huang, Horowitz, and Wei (2010)). We can then use the proposed method in this paper to determine linear and nonlinear components in the model. Under the conditions given in Huang, Horowitz, and Wei (2010) and those given in this paper, this two-step approach has the asymptotic oracle property even in $p > n$ settings. Further work is needed to evaluate the finite sample performance and to spell out the technical details of this two-step approach in $p > n$ settings.

The proposed semiparametric regression pursuit method extends the scope of the application of penalized methods from variable selection to structure estimation. We have focused on the proposed method in the context of semiparametric partially linear models. It can be extended to other models, such as the generalized partially linear and partially linear proportional hazards models (Huang (1999)). It would be interesting to generalize the results of this paper to these more complicated models.

Acknowledgements

J. Huang wishes to thank Professor Guang Cheng for sharing with us an advanced version of Zhang, Cheng, and Liu (2011) and Professor Cun-Hui Zhang for sharing his insights on the properties of the minimax concave penalty. We also thank an anonymous referee, an associate editor and the Editor for their helpful comments which led to considerable improvements in the paper. The research of Huang is partially supported by NIH grants R01CA120988, R01CA142774 and NSF grant DMS 0805670. The research of Ma is partially supported by NIH grants R01CA120988 and R01CA142774.

Appendix

Lemma A.1. If

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

then

$$\mathrm{P}\big(\max_{j\notin S_1} \|\tilde{\theta}_{nj} - \theta_{nj}\| > \theta_* - \gamma\lambda\big) \le O(1)\frac{(p-q)m_n}{\sqrt{n}(\theta_* - \gamma\lambda)}. \tag{A.1}$$

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

$$T_{nj} = (0_{m_n}, \ldots, 0_{m_n}, I_{m_n}, 0_{m_n}, \ldots, 0_{m_n}),$$

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 jth block. By the triangle inequality,

$$\|\tilde{\theta}_{nj} - \theta_{nj}\|_2 \le \|T_{nj}C_{(2)}^{-1}Z_{(2)}'Q\boldsymbol{\varepsilon}_n\|_2 + \|T_{nj}C_{(2)}^{-1}Z_{(2)}'Q\delta_n\|_2. \tag{A.2}$$

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

$$\begin{aligned}
\mathrm{E}\max_{j\notin S_1} \|T_{nj}C_{(2)}^{-1}Z_{(2)}'Q\boldsymbol{\varepsilon}_n\|_2 &\le n^{-1}\rho_{n1}^{-1}\mathrm{E}\|Z_{(2)}'Q\boldsymbol{\varepsilon}_n\|_2 \\
&= n^{-1/2}\rho_{n1}^{-1}\mathrm{E}\|n^{-1/2}Z_{(2)}'Q\boldsymbol{\varepsilon}_n\|_2 \\
&= n^{-1/2}\rho_{n1}^{-1}m_n^{-1/2}((p-q)m_n)^{1/2} \qquad \text{(A.3)} \\
&= O(1)(p-q)n^{-1/2}m_n. \qquad \text{(A.4)}
\end{aligned}$$

Thus

$$\mathrm{P}\big(\max_{j\notin S_1}\|T_{nj}C_{(2)}^{-1}Z_{(2)}'Q\boldsymbol{\varepsilon}_n\|\geq(\theta_*-\gamma)/2\big)\leq\frac{O(1)(p-q)m_n}{\sqrt{n}(\theta_*-\gamma\lambda)}.$$

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

$$\begin{aligned}\max_{j\notin S_1}\|T_{nj}C_{(2)}^{-1}Z_{(2)}'Q\delta_n\|_2&\leq\|nC_{(2)}^{-1}\|_2\cdot\|n^{-1}Z_{(2)}'Z_{(2)}\|_2^{1/2}\cdot\|n^{-1/2}\delta_n\|_2\\&=O_p(1)\rho_{n1}^{-1}\rho_{n2}^{1/2}(p-q)^{1/2}m_n^{-d}\\&=O_p(1)(p-q)^{1/2}m_n^{-(2d-1)/2}.\end{aligned}\tag{A.5}$$

Therefore, when

$$\frac{(p-q)m_n}{\sqrt{n}(\theta_*-\gamma\lambda)}\to0,$$

(A.1) holds.

Lemma A.2. If

$$\frac{1}{\lambda m_n^{(2d+1)/2}}\to0,$$

then

$$\mathrm{P}(n^{-1}\max_{j\in S_1}\|Z_j'H(\boldsymbol{\varepsilon}_n+\delta_n)\|>\lambda)\leq O(1)\frac{\{\log\{(q\vee1)m_n\}\}^{1/2}}{\lambda\sqrt{n}}.\tag{A.6}$$

Proof of Lemma A.2. Write

$$n^{-1}Z_j'H(\boldsymbol{\varepsilon}_n+\delta_n)=n^{-1}Z_j'H_n\boldsymbol{\varepsilon}_n+n^{-1}Z_j'H_n\delta_n.\tag{A.7}$$

By Lemma 2 of Huang, Horowitz, and Wei (2010),

$$\mathrm{E}\Big(\max_{j\in S_1}\|n^{-1/2}Z_j'H_n\boldsymbol{\varepsilon}_n\|_2\Big)\leq O(1)\big\{\log((p-|S_1|)m_n)\big\}^{1/2}.\tag{A.8}$$

Therefore,

$$\mathrm{P}\big(n^{-1}\max_{j\in S_1}\|Z_j'H_n\boldsymbol{\varepsilon}_n\|_2>\frac{\lambda}{2}\big)\leq O(1)\frac{\{\log(qm_n)\}^{1/2}}{\lambda\sqrt{n}}.\tag{A.9}$$

For the second term on the right hand side of (A.7),

$$\begin{aligned}n^{-1}\max_{j\in S_1}\|Z_j'H_n\delta_n\|_2&\leq n^{-1/2}\max_{j\in S_1}\|n^{-1}Z_j'Z_j\|_2^{1/2}\cdot\|H_n\|_2\cdot\|\delta_n\|_2\\&=O(1)\rho_{n2}^{1/2}(p-q)^{1/2}m_n^{-d}\\&=O(1)(p-q)^{1/2}m_n^{-(2d+1)/2}.\end{aligned}\tag{A.10}$$

Therefore, when

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

(A.6) follows from (A.9) and (A.10).

Proof of Theorem 1. Since $1/\sqrt{m_n}\gamma$ is less than the smallest eigenvalue of $Z'QZ/n$, $L(\cdot;\lambda,\gamma)$ in (2.8) is a convex function. By the Karush-Kuhn-Tucker conditions, a necessary and sufficient condition for $\hat{\theta}_n$ is

$$\begin{cases} Z_j'Q\big(y - Z\hat{\theta}_n\big) = n\dot{\rho}(\|\hat{\theta}_n\|;\lambda),\ \|\hat{\theta}_j\|_2 \neq 0, \\ \|Z_j'Q\big(y - Z\hat{\theta}_n\big)\|_2 \leq n\lambda, \qquad \|\hat{\theta}_{nj}\| = 0. \end{cases} \tag{A.11}$$

For $j \notin S_1$, if $\|\tilde{\theta}_{nj}\| \geq \gamma\lambda$, then $\dot{\rho}(\|\tilde{\theta}_{nj}\|;\lambda) = 0$. Thus $\tilde{\theta}_n$ satisfies (A.11) if also $\|Z_j'Q\big(y - Z\tilde{\theta}_n\big)\|_2 \leq n\lambda$ for $j \in S_1$. Therefore, $\hat{\theta}_n = \tilde{\theta}_n$ in the intersection of the events

$$\Omega_1(\lambda) = \Big\{ \min_{j\notin S_1} \|\tilde{\theta}_{nj}\| \geq \gamma\lambda \Big\} \text{ and } \Omega_2(\lambda) = \Big\{ \max_{j\in S_1} \|Z_j'Q(y - Z\tilde{\theta}_n)\| \leq n\lambda \Big\}. \tag{A.12}$$

Let $g_{0j}(x_j) = (g_{0j}(x_{1j}), \ldots, g_{0j}(x_{nj}))'$ and $\delta_n = \sum_{j\notin S_1} g_{0j}(x_j) - Z_{(2)}\theta_{n(2)}$. Let $C_{(2)} = Z_{(2)}'QZ_{(2)}$ and $H = Q - QZ_{(2)}(Z_{(2)}'QZ_{(2)})^{-1}Z_{(2)}'Q$. By (4.2),

$$\tilde{\theta}_{n(2)} - \theta_{n(2)} = C_{(2)}^{-1}Z_{(2)}'Q(\boldsymbol{\varepsilon}_n + \delta_n), \tag{A.13}$$

$$Z_j'Q(y - Z_{(2)}\tilde{\theta}_{n(2)}) = Z_j'H(\boldsymbol{\varepsilon}_n + \delta_n). \tag{A.14}$$

Recall that $\theta_* = \min_{j\in S_1} \|\theta_{nj}\|$. If $\|\tilde{\theta}_{nj} - \theta_{nj}\| \leq \theta_* - \gamma\lambda$, then $\min_{j\notin S_1} \|\tilde{\theta}_{nj}\| \geq \gamma\lambda$. Therefore,

$$1 - \mathrm{P}(\Omega_1(\lambda)) \leq \mathrm{P}(\max_{j\notin S_1} \|\tilde{\theta}_{nj} - \theta_{nj}\| > \theta_* - \gamma\lambda).$$

We also have

$$1 - \mathrm{P}(\Omega_2(\lambda)) \leq \mathrm{P}(n^{-1}\max_{j\in S_1} \|(Z_j'H(\boldsymbol{\varepsilon}_n + \delta_n)\| > \lambda).$$

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

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

$$\frac{1}{2n}\|Q(y - Z\hat{\theta}_n)\|_2^2 + \sum_{j=1}^{p} \rho_\gamma(\|\hat{\theta}_{nj}\|;\lambda) \leq \frac{1}{2n}\|Q(y - Z\theta_n)\|_2^2 + \sum_{j=1}^{p} \rho_\gamma(\|\theta_{nj}\|;\lambda). \tag{A.15}$$

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 $\|Q(y - Z\hat{\theta}_n)\|_2^2 = \|\nu_n\|_2^2 - 2\eta_n'\nu_n + \|\eta_n\|^2$. We can rewrite (A.15) as

$$\|\nu_n\|_2^2 - 2\eta_n'\nu_n \le 2n\sum_{j=1}^{p}\big(\rho_\gamma(\|\theta_{nj}\|;\lambda) - \rho_\gamma(\|\hat{\theta}_{nj}\|;\lambda)\big). \tag{A.16}$$

Since

$$\big|\rho_\gamma(\|\theta_{nj}\|;\lambda) - \rho_\gamma(\|\hat{\theta}_{nj}\|;\lambda)\big| \le \lambda\|\theta_{nj} - \hat{\theta}_{nj}\|, \tag{A.17}$$

combining (A.16) and (A.17), we get

$$\|\nu_n\|_2^2 - 2\eta_n'\nu_n \le 2n\lambda\sqrt{p}\|\hat{\theta}_n - \theta_n\|. \tag{A.18}$$

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

$$2|\eta_n'\nu_n| \le 2\|\eta_n^*\|_2 \cdot \|\nu_n\|_2 \le 2\|\eta_n^*\|_2^2 + \frac{1}{2}\|\nu_n\|_2^2. \tag{A.19}$$

From (A.18) and (A.19), we have

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

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

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

It follows that

$$\|\hat{\theta}_n - \theta_n\|_2^2 \le \frac{8\|\eta_n^*\|_2^2}{nc_{n*}} + \frac{4\lambda^2 p}{c_{n*}^2}. \tag{A.20}$$

Let $f_0(x_i) = \sum_{j=1}^{p} f_{0j}(x_{ij})$. Write

$$\eta_n = Q(\varepsilon_i + (\mu - \bar{y})\mathbf{1} + f(x_i) - Z\theta_n).$$

Since $|\mu - \bar{y}|^2 = O_p(n^{-1})$ and $\|f_{0j} - f_{nj}\|_\infty = O(m_n^{-d})$, we have

$$\|\eta_n^*\|_2^2 \le 2\|\boldsymbol{\varepsilon}_n^*\|_2^2 + O_p(1) + O(npm_n^{-2d}), \tag{A.21}$$

where $\boldsymbol{\varepsilon}_n^*$ is the projection of $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)'$ to the span of QZ. We have

$$\|\boldsymbol{\varepsilon}_n^*\|_2^2 = \|(Z'QZ)^{-1/2}Z'Q\boldsymbol{\varepsilon}_n\|_2^2 \le O_p(pm_n) \tag{A.22}$$

Combining (A.20), (A.21), and (A.22), we get

$$\|\hat{\theta}_n - \theta_n\|_2^2 \le O_p\Big(\frac{pm_n}{nc_{n*}}\Big) + O_p\Big(\frac{1}{nc_{n*}}\Big) + O\Big(\frac{d_{n2}m_n^{-2d}}{c_{n*}}\Big) + \frac{4p\lambda^2}{c_{n*}^2}.$$

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

$$\|\hat{\theta}_n - \theta_n\|_2^2 \le O_p\Big(\frac{pm_n^2}{n}\Big) + O_p\Big(\frac{m_n}{n}\Big) + O\Big(\frac{1}{m_n^{2d-1}}\Big) + O(m_n^2\lambda^2).$$

Now the result follows from the properties of polynomial splines (Schumaker (1981)).

Proof of Theorem 3. Let $\tilde{\theta}_n$ be the oracle estimator defined in (4.1), and let

$$\tilde{g}_{nj}(x) = 0, j \in S_1 \quad \text{and } \tilde{g}_{nj}(x) = \sum_{k=1}^{m_n} \tilde{\theta}_{jk}\psi_{jk}(x), j \in S_2,$$
$$\tilde{f}_{nj}(x) = \tilde{\beta}_j x + \tilde{g}_{nj}(x), \ j \in \hat{S}_2.$$

Write $\tilde{f}_{nj}(x_j) = (\tilde{f}_{nj}(x_{1j}), \ldots, \tilde{f}_{nj}(x_{nj}))'$. The estimator of the coefficients of the linear components is

$$\tilde{\beta}_{n1} = (X_{(1)}'X_{(1)})^{-1}X_{(1)}'(y - \sum_{j\in S_2} \tilde{f}_{nj}(x_j)).$$

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

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

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

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

This completes the proof of Theorem 3.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

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.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16**, 136-146.

Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81**, 310-320.

Friedman, J., Hastie, Hoefling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **35**, 302-332.

Fu, W. J. (1998). Penalized regressions: the bridge versus the LASSO. *J. Comp. Graph. Statist.* **7**, 397-416.

Härdle, W., Liang, H. and Gao, J. (2000). *Partially Linear Models*. Physica-Verlag, Heidelberg.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall.

Heckman, N. (1986). Spline smoothing in partly linear model. J. Roy. Statist. Soc. Ser. B **48**, 244-248.

Huang, J. (1996). Efficient estimation for the Cox model with interval censoring. *Ann. Statist.* **24**, 540-568.

Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27**, 1536-1563.

Huang, J., Horowitz, J. L. and Wei, F. R. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38**, 2282-2313.

Mazumder, R., Friedman, J. and Hastie, T. (2009). *SparseNet*: Coordinate descent with non-convex penalties. *J. Amer. Statist. Assoc.* **106**, 1125-1138.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica* **56**, 931-954.

Schumaker, L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13**, 970-983.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689-705.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109**, 475-494.

Wahba, G. (1984). Partial spline models for the semiparametric estimation of functions of several variables. *Analyses for Time Series, Japan-US Joint Seminar*, 319-329. Institute of Statistical Mathematics, Tokyo.

Willems, J. P., Saunders, J. T., Hunt, D. E. and Schorling, J. B. (1997). Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. *Southern Med. J.* **90**, 814-820.

Wu, T. and Lange, K. (2007). Coordinate descent procedures for lasso penalized regression. *Ann. Appl. Statist.* **2**, 224-244.

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

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

Zhang, H. H., Cheng, G. and Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *J. Amer. Statist. Assoc.* **106**, 1099-1112.

Zou, H. (2006). The adaptive Lasso and its oracle properties.*J. Amer. Statist. Assoc.* **101**, 1418-1429.

Department of Statistics and Actuarial Science and Department of Biostatistics, University of Iowa, Iowa City, Iowa 52242, USA.

E-mail: jian-huang@uiowa.edu

Department of Mathematics, University of West Georgia, Carrollton, Georgia 30118, USA.
E-mail: fwei@westga.edu

Division of Biostatistics, Department of Epidemiology and Public Health, Yale University, New Haven, Connecticut 06520, USA.
E-mail: shuangge.ma@yale.edu

(Received December 2010; accepted September 2011)