Statistica Sinica **25** (2015), 591-607
doi:http://dx.doi.org/10.5705/ss.2013.158

SEPARATION OF COVARIATES INTO NONPARAMETRIC AND PARAMETRIC PARTS IN HIGH-DIMENSIONAL PARTIALLY LINEAR ADDITIVE MODELS

Heng Lian, Hua Liang and David Ruppert

University of New South Wales, George Washington University and Cornell University

Abstract: Determining which covariates enter the linear part of a partially linear additive model is always challenging. It is more serious when the number of covariates diverges with the sample size. In this paper, we propose a double penalization based procedure to distinguish covariates that enter the nonparametric and parametric parts and to identify insignificant covariates simultaneously for the "large p small n" setting. The procedure is shown to be consistent for model structure identification, it can identify zero, linear, and nonlinear components correctly. The resulting estimators of the linear coefficients are shown to be asymptotically normal. We discuss how to choose the penalty parameters and provide theoretical justification. We conduct extensive simulation experiments to evaluate the numerical performance of the proposed methods and analyze a gene data set for an illustration.

Key words and phrases: Adaptive LASSO, curse of dimensionality, oracle property, penalized likelihood, polynomial splines, structure identification consistency.

1. Introduction

Consider the additive model (Hastie and Tibshirani (1990)):

$$Y = \mu + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{1.1}$$

where Y is a scalar response, $X = (X_1, \ldots, X_p)^T$ contains p covariates, μ is the intercept, f_j are unknown univariate functions. Let $(Y_i, X_i), i = 1, \ldots, n$ be independent and identically distributed (i.i.d.) as (Y, X).

Additional efforts have been made to further simplify model (1.1) and to balance the interpretability of linear models and flexibility of additive models. As a result, the partially linear additive model, a more parsimonious special case of (1.1), has been proposed and studied (Opsomer and Ruppert (1999); Liu, Wang, and Liang (2011)). If the choice of linear components is correctly specified, then

the biases in the estimation of these components are avoided and root-n convergence rates can be obtained for the linear coefficients. However, such prior knowledge is rarely available, especially when the number of covariates is large. Thus, determining which functions are linear is critical prior to use of (1.1). In this paper, we propose a penalization procedure for simultaneously identifying parametric components in an additive model and removing insignificant predictors when the numbers of the covariates can diverge with the sample size. Thus, it is not necessary to pre-determine the parametric components, and we show that the resulting estimator possesses the oracle property in the sense that it estimates as well as when zero components and parametric components are known a priori.

Penalization-based methods are traditionally used only for variable selection. Substantial progress has been made on linear regression when p is large, in particular based on lasso (Tibshirani (1996); Zou (2006); Zhao and Yu (2006); Huang, Ma and Zhang (2008); Zhang and Huang (2008); Bickel, Ritov and Tsybakov (2009)), which used an l_1 penalty to encourage shrinkage to zero. For additive models, several independent works (Ravikumar et al. (2009); Meier, Van de Geer, and Buhlmann (2009); Huang, Horowitz and Wei (2010)) have shown that the sparse additive models can be fitted successfully to various data sets with large dimensions. Given the success of sparse additive models in detecting insignificant predictors, it is highly desirable to identify the parametric components in a consistent framework, and the contributions of Zhang, Cheng, and Liu (2011) and Huang, Wei and Ma (2012) are discussed in Section 1.1.

The paper is organized as follows. After discussion of related works in the Section 2.1, we propose our doubly penalized estimation method and consider its asymptotic properties in Section 2.2. We focus on the adaptive lasso penalty only but note that other penalties such as smoothly clipped absolute deviation (SCD, Fan and Li (2001)) and minimax concave penalty (MCP, Zhang (2010)) could also be applied. The initial lasso estimation is discussed in Section 2.3. In Section 2.4, we adopt an extended BIC for tuning parameters selection (Chen and Chen (2008)) in the semiparametric setting and prove its consistency. The method is illustrated with extensive Monte Carlo simulations and an analysis of a data set in Section 3. The proofs for the main results are deferred to the Supplementary Appendix available online. We have placed R code for our implementation at `http://www.ntu.edu.sg/home/henglian/PLAMcode.htm`.

2. Penalized Estimation with Polynomial Splines

2.1. Related proposals

One possibility to simplify model (1.1) to a partial linear additive model is to fit an additive model (1.1) first, and then test component by component (Chen, Liang, and Wang (2011)). Another is to manipulate recursively the

following recipe (Liu, Wang, and Liang (2011)). Put all the continuous covariates in the nonparametric part and the discrete covariates in the parametric part. If the estimation results show that some of the continuous covariate effects can be described by certain parametric forms such as a linear form, then a new model can be fitted with those continuous covariate effects moved to the parametric part. Such approaches seem cumbersome in high-dimensional cases.

Our method is based on double penalization: one penalty function is used to identify zero components, and a second is used to identify parametric components. Double penalization strategies have been used before for other purposes, for example, in elastic net (Zou and Hastie (2005)), fused lasso (Tibshirani et al. (2005)), sparse group lasso (Peng et al. (2010)), and adaptive elastic-net lasso (Zou and Zhang (2009)). The idea behind our method is similar to those of Zhang, Cheng, and Liu (2011) and Huang, Wei and Ma (2012).

In Zhang, Cheng, and Liu (2011), starting from (1.1), the authors assume $f_j \in \mathcal{H}_j$ for some reproducing kernel Hilbert space (RKHS) $\mathcal{H}_j$ which admits the orthogonal decomposition $\mathcal{H}_j = \{1\} \oplus \mathcal{H}_{0j} \oplus \mathcal{H}_{1j}$, where $\{1\}$ is the space of constant functions and $\mathcal{H}_{0j}$ is a subspace of the space linear function (orthogonal to $\{1\}$). Let $\mathcal{P}_{0j}$ and $\mathcal{P}_{1j}$ be the orthogonal projection onto $\mathcal{H}_{0j}$ and $\mathcal{H}_{1j}$ respectively. Zhang, Cheng, and Liu (2011) propose to solve the problem

$$\min_{f_j \in \mathcal{H}_{0j} \cup \mathcal{H}_{1j}, \mu} \frac{1}{2} \sum_{i=1}^{n} (Y_i - \mu - \sum_{j=1}^{p} f_j(X_{ij}))^2 + n\lambda_1 \sum_{j=1}^{p} w_{0j} \|\mathcal{P}_{0j} f_j\|_{\mathcal{H}_j} + n\lambda_2 \sum_{j=1}^{p} w_{1j} \|\mathcal{P}_{1j} f_j\|_{\mathcal{H}_j},$$

where $\|.\|_{\mathcal{H}_j}$ is the RKHS norm, λ_1, λ_2 are regularization parameters, and w_{0j} and w_{1j} are appropriate weights. The idea is that if $\|\mathcal{P}_{1j} f_j\|_{\mathcal{H}_j} = 0$, then g_j is a linear function, and if $\|\mathcal{P}_{1j} f_j\|_{\mathcal{H}_j} = \|\mathcal{P}_{0j} f_j\|_{\mathcal{H}_j} = 0$, then $f_j = 0$.

Huang, Wei and Ma (2012) directly write $f_j(x) = \beta_{0j} + \beta_j x + g_j(x)$ with a series expansion $g_j(x) \approx \sum_{k=1}^{K} \theta_{jk} \phi_k(x)$ for some basis functions $\phi_1, \ldots, \phi_K$. Since f_j is linear if $\theta_j = (\theta_{j1}, \ldots, \theta_{jK})^T$ is zero, they proposed the penalized criterion

$$\min_{\theta_{jk}, \beta_j, \mu} \frac{1}{2} \sum_{i=1}^{n} (Y_i - \mu - \sum_{j=1}^{p} X_{ij} \beta_j - \sum_{j=1}^{p} \sum_{k=1}^{K} \theta_{jk} \phi_k(X_{ij}))^2 + np(\|\theta_j\|),$$

where $p(\cdot)$ is some penalty function (with some tuning parameters that we do not explicitly write down). In particular they used the group minimax concave penalty (Zhang (2010)). Although they only used one penalty and thus do not perform variable selection, it seems relatively easy to extend their results to the case with two penalties.

Our work differs from these papers in several ways: they focused only on fixed dimensional problems, while we consider a more challenging high-dimensional setting; we use an approach that directly shrinks the second derivative of the component function to zero; we construct a "regularized oracle estimator" and resort to results from convex analysis to overcome some theoretical difficulties due to high dimensionality.

2.2. Estimation procedure

Since the linear components are not pre-determined, the starting point of our analysis is the additive model (1.1). We assume the distribution of X_j is compactly supported and so, without loss of generality, supported on $[0,1]$. We also impose the condition $Ef_j(X_j) = 0$ for identifiability. We use polynomial splines to approximate the components. Let $\tau_0 = 0 < \tau_1 < \cdots < \tau_{K'} < 1 = \tau_{K'+1}$ be a partition of $[0,1]$ into subintervals $[\tau_k, \tau_{k+1}), k = 0, \ldots, K'$ with K' internal knots. We restrict attention to equally spaced knots, although a data-driven choice could be considered. A polynomial spline of order q is a function whose restriction to each subinterval is a polynomial of degree $q-1$ and which is globally $q-2$ times continuously differentiable on $[0,1]$. The collection of splines with a fixed sequence of knots has a normalized B-spline basis $\{B_1(x), \ldots, B_{\tilde{K}}(x)\}$ with $\tilde{K} = K' + q$. Because of the centering constraint $Ef_j(X_j) = 0$, we focus on the subspace of spline functions $S_j^0 := \{s : s = \sum_{k=1}^{\tilde{K}} b_{jk} B_k(x), \sum_{i=1}^n s(X_{ij}) = 0\}$ with basis $\{B_{jk}(x) = B_k(x) - \sum_{i=1}^n B_k(X_{ij})/n, k = 1, \ldots, K = \tilde{K} - 1\}$ (the subspace is $K = (\tilde{K} - 1)$-dimensional due to the constraint). Using spline expansions, we can approximate the components by $f_j(x) \approx \sum_k b_{jk} B_{jk}(x)$. It is possible to specify different values of K for each component but we assume they are the same for notational simplicity.

We are interested in a sparse model, in which many components f_j are zero, some components are linear, and the remaining ones are nonlinear. Without loss of generality, we assume the first p_1 components are nonlinear and should be modeled nonparametrically, the next p_2 components are linear, and all the rest are zero. Let $s = p_1 + p_2 \le p$ denote the total number of nonzero components. We assume the number of nonzero components is bounded and does not diverge with n.

Let $f_{0j}, 1 \le j \le p$, be the true functions and $\beta_0 = (\beta_{0,p_1+1}, \ldots, \beta_{0s})^T$ be the true coefficients in the linear components. We propose a penalized least squares

estimation procedure to automatically identify different types of components:

$$(\hat{\mu}, \hat{b}) = \arg\min_{\mu,b} \frac{1}{2} \sum_i \Big\{ Y_i - \mu - \sum_{j=1}^{p} \sum_{k=1}^{K} b_{jk} B_{jk}(X_{ij}) \Big\}^2 + n\lambda_1 \sum_{j=1}^{p} w_{1j} ||b_j||_{A_j} + n\lambda_2 \sum_{j=1}^{p} w_{2j} ||b_j||_{D_j}, \tag{2.1}$$

where λ_1, λ_2 are regularization parameters, $w_1 = (w_{11}, \ldots, w_{1p})$ and $w_2 = (w_{21}, \ldots, w_{2p})$ are vectors of weights that need to be appropriately chosen in order to achieve consistency in model selection. For now we assume these are given and leave their choices to the next subsection. We allow these weights to be random as is the case for weights derived from an initial lasso estimator (Section 2.3). A_j and D_j are two $K \times K$ matrices, $\|b_j\|_{A_j} = (b_j^T A_j b_j)^{1/2}$, and $\|b_j\|_{D_j} = (b_j^T D_j b_j)^{1/2}$. There is some flexibility in choosing A_j and D_j but one requirement is that $\|b_j\|_{A_j} = 0$ if and only if $\sum_k b_{jk} B_{jk}(x) \equiv 0$ and $\|b_j\|_{D_j} = 0$ if and only if $\sum_k b_{jk} B_{jk}(x)$ is a linear function, so that the two penalties can be used to identify zero and linear components, respectively. One natural choice is $A_j = \{\int_0^1 B_{jk}(x) B_{jk'}(x) dx\}_{k,k'=1}^K$ and $D_j = \{\int_0^1 B''_{jk}(x) B''_{jk'}(x) dx\}_{k,k'=1}^K$ (in this case all D_j's are the same by our construction of B_{jk}) so that $\|b_j\|_{A_j} = \|\sum_k b_{jk} B_{jk}\|$ and $\|b_j\|_{D_j} = \|\sum_k b_{jk} B''_{jk}\|$. Thus, both the estimated component and its second derivative are shrunk towards zero, as desired. Alternatively, we might set $A_j = I$, the identity matrix. Once we obtain estimates of b_{jk}, we estimate f_j using $\hat{f}_j = \sum_k \hat{b}_{jk} B_{jk}$.

Since μ in (2.1) is not penalized and B_{jk} is appropriately centered, it is straightforward to show $\hat{\mu} = \bar{Y} = \sum_i Y_i / n$. Using the notation

$$Z_j = \begin{pmatrix} B_{j1}(X_{1j}) & B_{j2}(X_{1j}) & \cdots & B_{jK}(X_{1j}) \\ \vdots & \vdots & \ddots & \vdots \\ B_{j1}(X_{nj}) & B_{j2}(X_{nj}) & \cdots & B_{jK}(X_{nj}) \end{pmatrix}_{n \times K},$$

$Z = (Z_1, \ldots, Z_p)$, and $Y = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})^T$, (2.1) can be written in matrix form as

$$\min_b \frac{1}{2} ||Y - Zb||^2 + n\lambda_1 \sum_{j=1}^{p} w_{1j} ||b_j||_{A_j} + n\lambda_2 \sum_{j=1}^{p} w_{2j} ||b_j||_{D_j}. \tag{2.2}$$

We now consider the asymptotic properties of the solution to (2.2). Proof of the folowing results are in the Supplementary Appendix.

Proposition 1. *Assume D_j satisfies the requirement that $\|b\|_{D_j} = 0$ if and only if $b^T B_j(x)$ is a linear function, where $B_j(x) = (B_{j1}(x), \ldots, B_{jK}(x))^T$. Fix $j \in \{p_1 + 1, \ldots, s\}$. For any $b_j \in R^K$, the following conditions are equivalent:*

(i) $\|b_j\|_{D_j} = 0$.
(ii) *There exists a unique constant* $a_j \in R$ *such that* $b_j^T B_j(x) \equiv a_j(x - \bar{X}_j)$ *where* $\bar{X}_j = \sum_{i=1}^n X_{ij}/n$. *Such a mapping from* R^K *to* R *is linear, one-one and onto, with* $\|b_j\|^2/K \sim a_j^2$.
(iii) b_j *is a constant multiple of* ξ_j, *where* $\xi_j \in R^K$ *satisfies* $\xi_j^T B_j(x) \equiv x - \bar{X}_j$.

Here $a_n \sim b_n$ means $0 < c \le a_n/b_n \le C < \infty$ for some constants c and C and all n.

Theorem 1. *Under Assumptions* (c1)–(c7) *in the Supplementary Appendix and assuming that* $K \log K/n \to 0$ *and* $K \to \infty$, *with probability approaching* 1, $\|\hat{b}_j\|_{D_j} = 0$ *for* $p_1 + 1 \le j \le s$ *and* $\hat{b}_j = 0$ *for* $s + 1 \le j \le p$.

Now we study the convergence rates of the nonzero components. Let $w_1^0 = (w_{11}, \ldots, w_{1s})$ and $w_2^0 = (w_{21}, \ldots, w_{2p_1})$. Here w_1^0, as a subvector of w_1, contains the weights associated with nonzero additive components, while w_2^0, a subvector, contains the weights in w_2 associated with nonlinear additive components. These weights are typically smaller in magnitude so that the corresponding components will not be shrunk to zero or to linear functions.

Theorem 2 (Convergence rates)**.** *Under the conditions of Theorem* 1, *the estimator obtained from* (2.2) *satisfies*

$$\sum_{j=1}^{s} ||\hat{f}_j - f_{0j}||^2 = O_p\left(\frac{K}{n} + \frac{1}{K^{2d}} + (\lambda_1^2||w_1^0||^2 + \lambda_2^2||w_2^0||^2)K\right), \tag{2.3}$$

where d *is the smoothness parameter for the component functions.*

For the parametric components, since $\|\hat{b}_j\|_{D_j} = 0$ *for* $p_1 + 1, \ldots, s$ *by Theorem* 1, *we effectively get estimates* $\hat{\beta}_j$ *of the true slopes* β_{0j}. *If* (c8) *holds and* $\sqrt{n}/K^{2d} \to 0$,

$$\sum_{j=p_1+1}^{s} (\hat{\beta}_j - \beta_{0j})^2 = O_p\left(\frac{1}{n} + (\lambda_1^2||w_1^0||^2 + \lambda_2^2||w_2^0||^2)K\right). \tag{2.4}$$

Under slightly stronger assumptions, the estimator for the parametric components can be shown to be asymptotically normal.

Theorem 3 (Asymptotic normality)**.** *Under the assumptions of Theorem* 2, *if* $\sqrt{nK}(\lambda_1\|w_1^0\| + \lambda_2\|w_2^0\|) = o_p(1)$

$$\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, \sigma^2\Xi^{-1}) \text{ in distribution},$$

with Ξ *the* $p_2 \times p_2$ *matrix at* (c8) *of the Supplementary Appendix.*

2.3. Initial estimator for determination of the weights

The adaptive group lasso penalty in (2.2) involves the weights w_1 and w_2. The weights w_{1j} are best if large for zero components and small for nonzero ones, and similarly the w_{2j} are best if large for linear components and small for nonparametric ones. There is some flexibility in specifying these weights based on a certain initial estimator. We mention that the group lasso estimator in Huang, Horowitz and Wei (2010) can be used; it is obtained from

$$\tilde{b} = \arg\min_b \frac{1}{2}\|Y - Zb\|^2 + n\lambda_0 \sum_{j=1}^{p} \|b_j\|_{A_j}.$$

Under certain assumptions, they showed that, if $\lambda_0 = C\sqrt{\log(pK)/n}$ for a sufficiently large constant C, then the number of estimated nonzero components is bounded by Ms for some constant M, and that $\sum_{j=1}^{p} \|\tilde{b}_j - b_{0j}\|_2^2 = O_p(K^2 \log(pK)/n + 1/K^{2d-1})$, where b_{0j} is any vector that satisfies $\|f_{0j} - b_{0j}^T B_j\| = O(K^{-d})$ for $j \le p_1$ and $f_{0j} = b_{0j}^T B_j$ for $j > p_1$. These authors only considered the case with A_j as the identity matrix, but the proof applies without change for any positive definite A_j whose eigenvalues are bounded and bounded away from zero.

Using this initial estimator, we can then set $w_{1j} = 1/\|\tilde{b}_j\|_{A_j}$ and $w_{2j} = 1/\|\tilde{b}_j\|_{D_j}$. Assuming $K \sim n^{1/(2d+1)}$ (this is the optimal choice that balances bias and variance term in the rates (2.3)), and $\log p = o(n^{2d/(2d+1)})$, we have $K^2 \log(pK)/n + 1/K^{2d-1} = o(K)$, and thus $\|\tilde{b}_j\|_{A_j} \ge C\sqrt{K}, 1 \le j \le s$ and $\|\tilde{b}_j\|_{D_j} \ge C\sqrt{K}$, $1 \le j \le p_1$, for some constant $C > 0$, under assumption (c9) in the Supplementary Appendix. This in turn implies $\|w_1^0\| = O_p(1/\sqrt{K})$ and $\|w_2^0\| = O_p(1/\sqrt{K})$. If we choose

$$\lambda_1 = \lambda_2 = O\Big(\sqrt{\frac{K}{n}}\Big), \tag{2.5}$$

the final term in the convergence rate (2.3) is at most the same order of the first term and can be ignored. Furthermore, based on the convergence rates for the initial estimator $\tilde{b}$ stated above, after some simple algebraic calculations, we see that assumption (c7) is satisfied if

$$\lambda_1 = \lambda_2 >> \frac{\sqrt{K}\log(pK) + K\sqrt{\log(pK)}}{n}. \tag{2.6}$$

For λ_1, λ_2 to satisfy both (2.5) and (2.6), we require that

$$\frac{\sqrt{K}\log(pK) + K\sqrt{\log(pK)}}{n} = o\left(\sqrt{\frac{K}{n}}\right),$$

which is the same as $p = o(\exp\{n^{1/2}\})$, the largest dimensionality allowed.

2.4. Tuning parameters selection

In practice, we need to choose some parameters including the spline order q, the number of basis K, as well as the regularization parameters λ_0, λ_1, and λ_2. We fixed $q = 4$ (cubic splines) in all our numerical results. When computing the initial group lasso estimator and the doubly penalized adaptive group lasso estimator, we fixed $K = 6$. This strategy is similar to that commonly used in functional smoothing/functional data analysis literature where the number of knots is chosen to be sufficiently large so that approximation error is small, and the overfitting can be effectively controlled by the penalization terms (see for example Section 5.5 of Ruppert, Wand, and Carroll (2003) or Chapter 5 of Ramsay and Silverman (2005)). The same strategy was adopted by Huang, Horowitz and Wei (2010) for high-dimensional additive models. In the simulation studies, we also computed the oracle estimator which is the minimizer of (A.2) in the Supplementary Appendix, with the penalty terms removed. In that case we used 10-fold cross-validation to choose K.

The choice of λ_1 and λ_2 in (2.2) is critical for the performance of the estimators. In our high-dimensional context, we adopt the extended Bayesian information criterion (eBIC) of Chen and Chen (2008) that was developed for parametric models. More specifically, we simultaneously select λ_1 and λ_2 based on the value of

$$\log\Big(\frac{1}{n}\|Y - Z\hat{b}_\lambda\|^2\Big) + d_1\frac{\log(n/K)}{n/K} + d_2\frac{\log n}{n} + \frac{d_1K + d_2}{n}\log p, \tag{2.7}$$

where $\hat{b}_\lambda$ is the minimizer of (2.2) for given $\lambda = (\lambda_1, \lambda_2)$, d_1 is the number of estimated nonparametric components and d_2 is the number of estimated parametric components, both for the given λ. For the initial estimator we use a similar criterion,

$$\log\Big(\frac{1}{n}\|Y - Z\hat{b}_\lambda\|^2\Big) + d_1\frac{\log(n/K)}{n/K} + \frac{d_1}{n/K}\log p. \tag{2.8}$$

In (2.7) and (2.8), if the last term is omitted, we have the ordinary BIC. We also note a slight difference of the final term in the criterion from that of Chen and Chen (2008), for example in (2.8) we use p^{d_1} instead of $\binom{p}{d_1}$. This is because when $\binom{p}{d_1}$ is used, it penalizes models with about half of its components being nonzero most heavily (since $\binom{p}{d_1}$ is largest when $d_1 \approx p/2$). In particular, based on this penalty, a full model is as parsimonious as the null model, which is unnatural (although the other penalty term can penalize against a large d_1). See Chen and Chen (2008) for the motivation of their use of $\binom{p}{d_1}$. Related work proving consistency of BIC or its modifications in parametric or nonparametric models includes Wang, Li, and Tsai (2007); Wang and Xia (2009); Wang, Li, and Leng (2009).

Theorem 4. *For models with at most S (does not increase with n) nonzero components, $S \geq s$, then under the conditions that $K \sim n^{1/(2d+1)}$ and $K \log p/n \to 0$, in addition to those assumed in Theorems* 1 *and* 2*, the extended BIC* (2.7) *will correctly identify the nonzero components and the parametric components with probability approaching* 1.

We do not have theoretical performance guarantees for BIC or eBIC in the initial non-adaptive group lasso estimator. The choice of criterion is investigated in detail in Monte Carlo studies.

3. Numerical Examples

3.1. Simulations

The minimization problem (2.2) is solved by local quadratic approximation as adopted in Fan and Li (2001). Given the current estimate $b_j^{(0)}$, the penalty terms can be approximated by

$$\|b_j\|_{A_j} \approx \|b_j^{(0)}\|_{A_j} + \frac{1}{2}\frac{\|b_j\|_{A_j}^2 - \|b_j^{(0)}\|_{A_j}^2}{\|b_j^{(0)}\|_{A_j}},$$

$$\|b_j\|_{D_j} \approx \|b_j^{(0)}\|_{D_j} + \frac{1}{2}\frac{\|b_j\|_{D_j}^2 - \|b_j^{(0)}\|_{D_j}^2}{\|b_j^{(0)}\|_{D_j}}.$$

After removing some irrelevant terms, the optimization problem becomes quadratic in b and has a closed-form solution. During the iterations, as soon as some $\|b_j\|_{A_j}$ (respectively, $\|b_j\|_{D_j}$) drops below a certain threshold (10^{-6} in our implementation), the component is identified as a zero function (respectively, linear function).

We generated data from the model

$$Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i,$$

with $f_1(x) = 5\sin(2\pi x)$, $f_2(x) = 10x(1-x)$, $f_3(x) = 3x$, $f_4(x) = 2x$, $f_5(x) = -2x$, $f_j(x) = 0, j > 5$ and $\epsilon_i \sim N(0, \sigma^2)$. Thus in our simulations $p_1 = 2$ and $p_2 = 3$. To generate covariates, we first let X_{ij} be marginally standard normal with correlations given by $\mathrm{Cov}(X_{ij_1}, X_{ij_2}) = (1/2)^{|j_1 - j_2|}$, and then applied the cumulative distribution function of standard normal distribution to transform X_{ij} to be marginally uniform on $[0, 1]$. We performed simulations with $n = 50, 100, 200$, $p = 50, 100, 200$, and $\sigma = 0.2, 0.5$, resulting in a total of 18 scenarios. For each scenario, the reported numerical results are based on 50 simulated data sets. To save space, only the case $n = 100$ is shown below (6 scenarios) and the complete simulation results are presented in the supplementary material available online.

First we consider the model selection performance of the non-adaptive group lasso estimator and adaptive group lasso estimator, when BIC or eBIC is used for choosing the regularization parameters. As mentioned in Section 2.4, we fix $K = 6$ in all experiments below. Table 1 reports the model identified by various estimation procedures. The initial non-adaptive lasso estimators with a single penalty cannot identify linear components. If BIC is used, there is a large number of identified nonzero components (false positives) and if eBIC is used, it seems the penalty is too strong so that some nonzero components are missed (false negatives). Since zero coefficients in the initial estimator will stay zero in the adaptive group lasso estimator, these false negatives cannot be addressed by the subsequent estimator. On the other hand, if BIC is used in the initial estimator, although there are a large number of false positives, these will be corrected in the second step if eBIC is used, as seen from the table. Finally, if both steps adopt ordinary BIC, the number of false positive is still too large. In summary, the best performance in model selection is obtained when BIC is used for initial estimator and eBIC is used for doubly penalized adaptive estimator. Thus we only consider this combination of criteria in the following.

In Table 2, we present the root mean squared errors for the first 6 component functions (note f_6 is zero),

$$RMSE_j = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(\hat{f}_j(t_i) - f_j(t_i))^2},$$

evaluated on a fine grid $(t_1, \ldots, t_T)$ consisting of 500 equally spaced points on $[0, 1]$. We compare our estimator to two others, including an oracle estimator where the nonlinear and linear components are known before analysis, and a sparse additive model (SA) where only one penalty is used in both steps to identify nonzero components only. Comparing SA with our estimator, we see that for the nonparametric components (f_1 and f_2), RMSEs are similar. However, for the truly linear components, our doubly penalized estimators obviously have better performance than SA, with improvement on RMSE of 30–50%. To see how this efficiency in estimation can result in better prediction, we also generated n independent test observations in each scenario and the squared prediction errors are reported in a table in the supplementary material. Our estimator is seen there to produce smaller predictor errors than SA. We also considered linear estimators based on the adaptive lasso and the RMSE for the linear model is much larger, which is not surprising since some components are truly nonlinear in the simulations. Therefore, the results for linear models are not reported.

Finally, we investigate the estimation of the standard errors of the linear coefficients. Since after spline approximation, computationally the model is similar

Table 1. Model identification results with $n = 100$. The different rows in each scenario correspond to different estimators/parameter selection criteria. BIC: the initial estimator using BIC for regularization parameter selection. EBIC: the initial estimator using eBIC for parameter selection. BIC/BIC: the final estimator obtained by using BIC for parameter selection in the initial estimator, and also using BIC for the doubly penalized adaptive estimator in the second step. EBIC/EBIC and BIC/EBIC are interpreted similarly. #N: number of nonparametric components identified; #NT: number of nonparametric components identified that are truly nonparametric (or truly nonzero for the initial estimator). #L: number of linear components identified; #LT: number of linear components identified that are truly linear. The true number of nonparametric components is 2 and the true number of linear components is 3. The numbers in smaller font are the corresponding standard errors.

		#N	#NT	#L	#LT
$n = 100$	BIC	$32.86_{15.1563}$	5_{0}	0_{0}	0_{0}
$p = 50$	EBIC	$6.22_{2.8521}$	$4.36_{1.3667}$	0_{0}	0_{0}
σ=0.2	BIC/BIC	$2.6_{0.8571}$	2_{0}	$2.74_{1.2747}$	$2.4_{0.8571}$
	EBIC/EBIC	$1.94_{0.5115}$	$1.84_{0.3703}$	$2.48_{1.1110}$	$2.42_{1.0515}$
	BIC/EBIC	$2.06_{0.2399}$	2_{0}	$3.06_{0.5500}$	$2.94_{0.2399}$
$n = 100$	BIC	$43.46_{1.5281}$	5_{0}	0_{0}	0_{0}
$p = 50$	EBIC	$3.74_{2.2840}$	$3.3_{1.7871}$	0_{0}	0_{0}
$\sigma = 0.5$	BIC/BIC	$12.64_{12.5727}$	2_{0}	$2.8_{1.7261}$	$1.44_{1.3273}$
	EBIC/EBIC	$1.6_{0.5714}$	$1.56_{0.5014}$	$1.7_{1.3132}$	$1.7_{1.3132}$
	BIC/EBIC	$2.42_{0.5380}$	2_{0}	$3.32_{1.0583}$	$2.64_{0.4849}$
$n = 100$	BIC	$25.7_{19.1644}$	$4.9_{0.5803}$	0_{0}	0_{0}
$p = 100$	EBIC	$4.92_{3.0159}$	$3.78_{1.6817}$	0_{0}	0_{0}
$\sigma = 0.2$	BIC/BIC	$2.98_{1.4497}$	2_{0}	$2.68_{1.7076}$	$2.16_{1.1314}$
	EBIC/EBIC	$1.76_{0.5175}$	$1.72_{0.4536}$	$2_{1.3093}$	$2_{1.3093}$
	BIC/EBIC	$2.04_{0.2828}$	$1.98_{0.1414}$	$3.04_{0.7548}$	$2.86_{0.4953}$
$n = 100$	BIC	$25.2_{24.3788}$	$4.8_{0.8081}$	0_{0}	0_{0}
$p = 100$	EBIC	$3.84_{2.6447}$	$3.16_{1.6826}$	0_{0}	0_{0}
$\sigma = 0.5$	BIC/BIC	$4.2_{6.8512}$	$1.94_{0.2399}$	$2.62_{1.3536}$	$2.2_{1.1429}$
	EBIC/EBIC	$1.68_{0.7677}$	$1.52_{0.5047}$	$1.44_{1.2316}$	$1.42_{1.1968}$
	BIC/EBIC	$2.26_{0.5997}$	$1.96_{0.1979}$	$2.76_{0.8704}$	$2.58_{0.7309}$
$n = 100$	BIC	$13.42_{11.8754}$	$4.6_{1.1066}$	0_{0}	0_{0}
$p = 200$	EBIC	$3.8_{3.0034}$	$2.86_{1.7958}$	0_{0}	0_{0}
$\sigma = 0.2$	BIC/BIC	$2.88_{1.0230}$	2_{0}	$2.36_{1.4107}$	$2.1_{1.0738}$
	EBIC/EBIC	$1.58_{0.6728}$	$1.48_{0.5047}$	$1.3_{1.4178}$	$1.24_{1.3180}$
	BIC/EBIC	$2.18_{0.6289}$	$1.92_{0.2740}$	$2.52_{1.0349}$	$2.4_{0.9258}$
$n = 100$	BIC	$9.48_{9.1724}$	$4.14_{1.4709}$	0_{0}	0_{0}
$p = 200$	EBIC	$2.42_{1.7507}$	$2.1_{1.1995}$	0_{0}	0_{0}
$\sigma = 0.5$	BIC/BIC	$2.18_{1.3506}$	$1.68_{0.4712}$	$2.02_{1.5451}$	$1.72_{1.1787}$
	EBIC/EBIC	$1.28_{0.4536}$	$1.28_{0.4536}$	$0.82_{0.9624}$	$0.8_{0.9476}$
	BIC/EBIC	$1.96_{0.6987}$	$1.76_{0.4314}$	$2.44_{1.2644}$	$2.22_{1.0746}$

Table 2. Root mean squared errors for the first six components with $n = 100$. "Sparse Additive" denotes the estimator for the sparse additive model obtained when $\lambda_2 = 0$ in (2.2). The numbers in smaller font are the corresponding standard errors.

		Oracle	Our Estimator	Sparse Additive
$n = 100$	f_1	$0.3286_{0.01716}$	$0.3301_{0.01622}$	$0.3301_{0.01883}$
$p = 50$	f_2	$0.0761_{0.02542}$	$0.1184_{0.04923}$	$0.0883_{0.03043}$
$\sigma = 0.2$	f_3	$0.0319_{0.02216}$	$0.0361_{0.02643}$	$0.0800_{0.02745}$
	f_4	$0.0366_{0.02271}$	$0.0481_{0.03229}$	$0.0941_{0.04142}$
	f_5	$0.0361_{0.02702}$	$0.0432_{0.03751}$	$0.0929_{0.04113}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
$n = 100$	f_1	$0.3364_{0.01925}$	$0.3468_{0.03131}$	$0.3420_{0.02067}$
$p = 50$	f_2	$0.1186_{0.04531}$	$0.1753_{0.08853}$	$0.1541_{0.06361}$
$\sigma = 0.5$	f_3	$0.0527_{0.03669}$	$0.0645_{0.04389}$	$0.1601_{0.07305}$
	f_4	$0.0494_{0.04048}$	$0.0707_{0.05324}$	$0.1812_{0.06942}$
	f_5	$0.0463_{0.03850}$	$0.0634_{0.05495}$	$0.1733_{0.08265}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
$n = 100$	f_1	$0.3257_{0.01730}$	$0.3308_{0.02130}$	$0.3302_{0.02229}$
$p = 100$	f_2	$0.0732_{0.02724}$	$0.1162_{0.09691}$	$0.0919_{0.09643}$
$\sigma = 0.2$	f_3	$0.0346_{0.02900}$	$0.0520_{0.12083}$	$0.0981_{0.11482}$
	f_4	$0.0386_{0.02964}$	$0.0682_{0.10873}$	$0.1120_{0.10062}$
	f_5	$0.0347_{0.03369}$	$0.0513_{0.08559}$	$0.1008_{0.08169}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
$n = 100$	f_1	$0.3382_{0.02306}$	$0.3466_{0.02564}$	$0.3452_{0.03010}$
$p = 100$	f_2	$0.1206_{0.04630}$	$0.1881_{0.13073}$	$0.1723_{0.12995}$
$\sigma = 0.5$	f_3	$0.0437_{0.03314}$	$0.0855_{0.16233}$	$0.1633_{0.15331}$
	f_4	$0.0480_{0.03817}$	$0.1068_{0.13266}$	$0.1923_{0.14004}$
	f_5	$0.0487_{0.03855}$	$0.0868_{0.13342}$	$0.1615_{0.13449}$
	f_6	$0.0000_{0.00000}$	$0.0022_{0.01577}$	$0.0026_{0.01835}$
$n = 100$	f_1	$0.3243_{0.01170}$	$0.3385_{0.03798}$	$0.3384_{0.03830}$
$p = 200$	f_2	$0.0770_{0.02476}$	$0.1626_{0.17174}$	$0.1403_{0.17663}$
$\sigma = 0.2$	f_3	$0.0366_{0.03136}$	$0.1080_{0.22410}$	$0.1428_{0.21421}$
	f_4	$0.0366_{0.03087}$	$0.1186_{0.18542}$	$0.1574_{0.17160}$
	f_5	$0.0517_{0.03874}$	$0.4503_{0.19922}$	$0.5215_{0.13808}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
$n = 100$	f_1	$0.3365_{0.01910}$	$0.3712_{0.06386}$	$0.3733_{0.05478}$
$p = 200$	f_2	$0.1189_{0.03959}$	$0.2875_{0.25395}$	$0.2717_{0.24912}$
$\sigma = 0.5$	f_3	$0.0530_{0.03713}$	$0.1679_{0.26175}$	$0.2532_{0.23850}$
	f_4	$0.0503_{0.04106}$	$0.2073_{0.23334}$	$0.3016_{0.20832}$
	f_5	$0.0482_{0.03595}$	$0.1924_{0.23047}$	$0.2686_{0.20975}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$

to a linear model, it is easy to adapt the sandwich formula proposed in Fan and Li (2001) for estimation of standard errors. For each of the three linear coefficients in the true model, the sample standard deviation (SD) of the estimated

coefficients in the simulations (only using simulations in which the component is correctly identified as linear) is regarded as the true standard error. The estimated standard error (SE) averaged over repeated simulations, together with the sample standard deviation of the estimated standard errors, are calculated. We find that the sandwich formula works well only for large sample sizes. The results are reported in a table in the supplementary material. The unsatisfactory performance of standard error estimation based on asymptotic normality for sparse regression has been noted in the literature, for example in Chatterjee and Lahiri (2013). It would be interesting to investigate better standard error estimation in the future.

3.2. Data

Here our main purpose is to use data to demonstrate that the automatically identified semiparametric additive models can be more accurate in prediction than general additive models as well as linear models.

An interesting question in fundamental biological research is that whether gene expression pattern can be determined by the gene's upstream sequence. This problem was investigated in Meier, Van de Geer, and Buhlmann (2009) using an additive model to predict gene expression levels. Using a general additive model with only nonparametric components is a reasonable strategy when little is known about whether the genes have linear effects or not. However, one would wonder whether a partially linear additive model can make the estimation more efficient. Of course, the difficulty is that it is a priori unknown which genes should enter the linearly.

We applied our method to the ChIP-chip data of *Saccharomyces cerevisiae* reported in Lee et al. (2002) and also analyzed by Hong et al. (2005) for motif discovery. In contrast to these works, here our goal was to predict binding intensities based on the DNA sequence. There are a total of forty data sets, for genes targeted by forty different transcription factors (TF). Based on an ChIP-chip p-value of 0.001 as the cutoff, from 25 to 176 positive sequences (believed to contain binding sites for the TF according to ChIP-chip experimental results) were obtained by Hong et al. (2005) for different TFs. For each positive sequence i in a data set, motif scores $x_{ij}, j = 1, \ldots, p$, are available as covariates in the regression, obtained as follows. First the Gibbs sampling program AlignACE (Roth et al. (1998); Hughes et al. (2000)) was run to find at most 100 motifs with MAP score (MAP score is a metric for motif strength used by AlignACE) at least 10. Because of this constraint, in many data sets a much smaller number of motifs are found, resulting in dimension p ranging from 27 to 100 for different data sets. Then the motif scores were found for each sequence-motif pair as in

Table 3. Prediction errors for the motif regression data sets. 12 TFs are used in the analysis. The smallest prediction error among the three estimators are in boldface.

TF	#seq	# motif	our estimator	sparse additive	sparse linear
ABF1	176	99	**5.140**	5.331	5.587
CIN5	116	100	2.262	**2.142**	2.587
FHL1	124	100	**3.752**	3.876	4.497
FKH2	72	99	**2.773**	2.994	4.423
GCN4	56	100	**3.266**	3.271	4.529
MBP1	74	35	0.863	1.063	**0.715**
MCM1	58	87	1.540	1.856	**1.534**
NDD1	66	100	**1.819**	1.898	2.103
RAP1	127	100	**2.098**	2.102	3.066
REB1	89	75	0.839	**0.838**	1.043
SWI4	90	100	**3.296**	3.797	4.380
SWI6	65	27	**0.289**	0.327	0.307

Motif Regressor (Conlon et al. (2003)), which is a matching score representing the existence of motif in the sequence.

For this analysis, we only used a subset of the data sets with at least 50 positive sequences and for which the known true motif can be found by AlignACE, resulting in a total of 12 data sets that are listed in Table 3. The 2nd and 3rd columns of the table show the number of positive sequences and the number of motifs found respectively for each TF. For each data set there exists a large number of negative sequences which are believed to be not binding to the TF. We randomly selected from it the same number of negative sequences as the number of positive sequences and motif scores were also calculated on the negative sequences. Half of the sequences were used for training in regression and the rest were used for testing the prediction accuracy on the binding intensities. The prediction errors for our estimator, the sparse additive estimator and the sparse linear estimator are reported in Table 3 as the last three columns. It is seen that our estimator performs best among the three for 8 out of 12 data sets. Even for those data sets where our estimator is not the best, it can be seen that the differences are small and our estimator is never the worst of the three estimators. The good performance of our estimator can be attributed to its ability to automatically choose a model specification that represents a reasonable trade-off between efficiency and flexibility.

4. Conclusion and Discussion

We here proposed a data-based procedure for identifying the nonparametric and parametric components of semiparametric additive models. Based on a double penalization strategy, model identification is performed simultaneously with

estimation, which is not possible with existing methods. Our Monte Carlo studies and applications to a data set demonstrate that the sparse partially linear additive model (with automatically determined model structure) can be more efficient than sparse additive models and can improve predictions. In high-dimensional settings, we naturally expect that some components are linear for parsimony and some components are nonparametric for flexibility. The proposed framework is thus valuable for automatically determining the linear part of a semiparametric model.

The extended BIC was shown to work very well in our numerical examples for choosing appropriate tuning parameters. The consistency proof of eBIC requires the assumption that $K \sim n^{1/(2d+1)}$, but in practice we fixed K to be a more or less arbitrary integer. In a high-dimensional case, it is generally difficult to automatically determine K, or even prove the resulting K has the desired divergence rate. One possibility is to choose K based on some hold-out data as in cross-validation methods, but this search would increase the computational burden of the method.

In terms of computation, we used the local quadratic approximation approach of Fan and Li (2001). One shortcoming of this algorithm is that it cannot obtain exactly zero solutions. To address this problem, algorithms of the coordinate-descent type have been used for sparse estimators (Zou and Li (2008); Huang, Horowitz and Wei (2010); Huang, Wei and Ma (2012)). This is based on the observation that when all coefficients are fixed except those associated with one predictor, there is a closed-form solution to the optimization problem via thresholding. However, with two penalties, it seems not straightforward to extend this class of algorithms for our problem. How to obtain exact zero solutions for our problem in the future is an interesting problem.

It seems possible to extend the double penalization approach to generalized additive models in order to deal with different types of responses. It is also of interest to see how it works with quantile regression, which provides a more complete picture of the relationship between responses and predictors and is also more robust than mean regression.

Acknowledgements

The authors want to thank the AE and an anonymous reviewer for their insightful comments and suggestions that led to significant improvement of the manuscript. Liang's research was partially supported by NSF grants DMS-1007167 and DMS-1207444, and by Award Number 11228103, given by National Natural Science Foundation of China. Lian's research was supported by a Singapore MOE Tier 2 Grant. Ruppert's research was supported by NSF grant DMS-080597.

References

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector, *Ann. Statist.* **37**, 1705-1732.

Chatterjee, A. and Lahiri, S. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *Ann. Statist.* **41**, 1232-1259.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759-771.

Chen, R., Liang, H. and Wang, J. (2011). On determination of linear components in additive models. *J. Nonparametr. Stat.* **23**, 367-383.

Conlon, E. M., Liu, X. S., Lieb, J. D. and Liu, J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. USA* **100**, 3339-3344.

Ekeland, I. and Turnbull, T. (1983). *Infinite-Dimensional Optimization And Convexity*, University of Chicago Press.

Fan, J. Q. and Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

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

Hong, P., Liu, X. S., Zhou, Q., Lu, X., Liu, J. S. and Wong, W. H. (2005). A boosting approach for motif modeling using chip-chip data. *Bioinformatics* **21**, 2636-2643.

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

Huang, J., Ma, S. G. and Zhang, C. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statist. Sinica* **18**, 1603-1618.

Huang, J., Wei, F., and Ma, S. (2012). Semiparametric regression pursuit. *Statist. Sinica* **22**, 1403-1426.

Hughes, J. D., Estep, P. W., Tavazoie, S. and Church, G. M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae1. *J. Molecular Biology* **296**, 1205-1214.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M. and Simon, I. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science* **298**, 799-804.

Li, Q. (2000). Efficient estimation of additive partially linear models. *Internat. Econom. Rev.* **41**, 1073-1092.

Liu, X., Wang, L. and Liang, H. (2011). Variable selection and estimation for semiparametric additive partial linear models. *Statist. Sinica* **21**, 1225-1248.

Meier, L., Van de Geer, S., and Buhlmann, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37**, 3779-3821.

Opsomer, J. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *J. Comput. Graph. Statist.* **8**, 715-732.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R. and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Statist.* **4**, 53-77.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd edition. Springer, New York.

Ravikumar, P., Lafferty, H., Liu, H. and Wasserman, L. (2009). Sparse additive models. *J. Roy. Statist. Soc. Ser. B* **71**, 1009-1030.

Roth, F. P., Hughes, J. D., Estep, P. W. and Church, G. M. (1998). Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature Biotechnology* **16**, 939-945.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. Roy. Statist. Soc. Ser. B* **67**, 91-108.

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

Wang, H. S. and Xia, Y. C. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.* **104**, 747-757.

Wang, H. S., Li, B. and Leng, C. L. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. Roy. Statist. Soc. Ser. B* **71**, 671-683.

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

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

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.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541-2563.

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

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

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

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37**, 1733-1751.

School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, 2052, Australia.

E-mail: henglian@ntu.edu.sg

Department of Statistics, George Washington University, 801 22nd St. NW, Washington, D.C. 20052, U.S.A.

E-mail: hliang@gwu.edu

School of Operations, Research and Information Engineering, Cornell University, Ithaca, N.Y. 14853, U.S.A.

E-mail: dr24@cornell.edu

(Received June 2013; accepted May 2014)