Published in final edited form as:

J Nonparametr Stat. 2012 June 1; 24(2): 283–298. doi:10.1080/10485252.2012.661054.

Variable selection for semiparametric regression models with iterated penalization

Ying Dai* and Shuangge Ma

School of Public Health, Yale University, New Haven, CT, USA

Abstract

Semiparametric regression models with multiple covariates are commonly encountered. When there are covariates not associated with response variable, variable selection may lead to sparser models, more lucid interpretations and more accurate estimation. In this study, we adopt a sieve approach for the estimation of nonparametric covariate effects in semiparametric regression models. We adopt a two-step iterated penalization approach for variable selection. In the first step, a mixture of the Lasso and group Lasso penalties are employed to conduct the first-round variable selection and obtain the initial estimate. In the second step, a mixture of the weighted Lasso and weighted group Lasso penalties, with weights constructed using the initial estimate, are employed for variable selection. We show that the proposed iterated approach has the variable selection consistency property, even when number of unknown parameters diverges with sample size. Numerical studies, including simulation and analysis of a diabetes dataset, show satisfactory performance of the proposed approach.

Keywords

Iterated penalization; Variable selection; Semiparametric regression

1. Introduction

In regression analysis, semiparametric models are commonly encountered. Compared with parametric models, semiparametric models are more flexible and provide better description of subtle data structure. It is common that multiple covariates are measured, among which some are associated with response variable while others are not. A representative dataset is described in Section 4. With such data, it is desirable to conduct variable selection along with estimation. In semiparametric regression analysis, variable selection may lead to sparser models, more lucid interpretations and more accurate estimation.

Consider the regression model

$$Y = \alpha + \sum_{j=1}^{p_1} \beta_j X_j + \sum_{j=p_1+1}^p f_j(X_j) + \varepsilon. \quad (1)$$

Here Y is the response variable, α is the intercept, $(X_1, \dots, X_p)$ are the p -dimensional covariates, $(\beta_1, \dots, \beta_{p_1})$ are the p_1 parametric regression coefficients, f_j s are the non-parametric covariate effects, and ε is the random error. There are a total of p covariate

effects, among which p_1 are parametric, and the rest $p - p_1$ are nonparametric. For identifiability, assume that $E(f_j(X_j)) = 0$, $j = p_1 + 1, \dots, p$, where E denotes expectation. The estimation aspect of semiparametric regression has been well investigated. See for example Ruppert, Wand and Carroll (2003), Horowitz (2009) and many others. In this article, we focus on the scenario where some β_j s and/or f_j s are exactly equal to zero, indicating that the corresponding covariates have no effect on the response variable. Our goal is to conduct variable selection and identify covariates with nonzero effects.

With parametric models, a large number of approaches have been developed for variable selection. Among available approaches, penalization approaches have attracted extensive attention because of their satisfactory statistical and numerical properties. Available penalization approaches include Lasso, bridge, SCAD, MCP, elastic net and others (Tibshirani 1996; Huang et al. 2008; Fan and Li 2001; Zhang and Huang 2008; Zhang 2010). Under certain regularity conditions, several approaches, including adaptive Lasso (Zou 1996), SCAD (Fan and Li 2001), bridge (Huang et al. 2008) and MCP (Zhang 2010), have the much desired selection consistency property. Such a result may hold even when the number of covariates diverges with sample size. For some datasets, there exist grouping structures among covariates, where covariates within the same groups tend to be “related” and different groups tend to be “unrelated”. With such data, the group counterparts of the aforementioned penalization approaches have been developed and can respect the grouping structures in variable selection.

Variable selection with semiparametric and nonparametric regression models has also been investigated in recent studies. A family of methods are based on the hypothesis testing $H_0: f_j = 0$ (Wand 2003). In practice, such methods make basis expansions of f_j s, transform the problem into a multi-dimensional parametric hypothesis testing problem and have been shown to have satisfactory numerical performance. However, as a large number of bootstraps are needed to generate the null distribution, they may suffer a high computational cost particularly with multiple nonparametric covariate effects. For example, the analysis of a single dataset with 200 samples and four nonparametric covariate effects may take over an hour. In addition, the asymptotic properties of such methods have not been rigorously established. Liang and Li (2008) propose using the SCAD penalty for variable selection with parametric covariate effects and a semiparametric generalized likelihood ratio test for variable selection with nonparametric covariate effects. Despite its satisfactory statistical properties, the SCAD penalized estimation is difficult to compute. In Xie and Huang (2009) and several other studies, semiparametric regression models with multi- and high-dimensional covariates are studied. However, the focus has been on variable selection with parametric covariate effects while properly adjusting for a fixed number of nonparametric covariate effects. In a recent study, Du et al. (2010) describe a Kullback-Leibler distance-based approach for variable selection with nonparametric covariate effects. The asymptotic property of this approach has not been established. In addition, in practice, it has been applied only to data with a small number of nonparametric covariate effects.

With semiparametric model (1), we adopt a sieve approach and make basis expansion for estimation of nonparametric covariate effects. We adopt penalization for variable selection. Among the available penalization approaches, the simplest one is the Lasso (Tibshirani 1996) where the penalty is the L_1 norm of regression coefficients. For parametric covariate effects, variable selection can be straightforwardly achieved using Lasso; For nonparametric covariate effects, with multiple basis functions, one covariate effect corresponds to a *group* of multiple regression coefficients. The group Lasso approach (Yuan and Lin 2006) can then be adopted for variable selection. As shown in Zhao and Yu (2006), Zhang and Huang (2008) as well as Section 3, a drawback of Lasso-based approaches is that they are in general not variable selection consistent unless strong irrepresentable conditions are

satisfied. To achieve variable selection consistency, an iterated Lasso-based penalization method is adopted in this article.

This study differs from previous studies along the following directions. Compared with variable selection studies on parametric models, a semiparametric model is adopted and can better describe the subtle nonlinear relationship in data. Compared with the hypothesis testing-based approaches, the proposed approach has a lower computational cost. For example, the analysis of a single dataset with 200 samples and 500 covariates takes less than five minutes on a desktop PC. In addition, its asymptotic properties can be rigorously established. Compared with the SCAD penalization approaches, the proposed approach is built on the Lasso penalties and easier to compute. Compared with the Kullback-Leibler distance-based approach, the proposed approach has more lucid asymptotic properties and can more easily accommodate multiple nonparametric covariate effects. The rest of the article is organized as follows. In Section 2, we describe the iterated penalization approach. We also describe an effective coordinate descent computational algorithm. In Section 3, we establish the selection consistency property. Numerical studies, including simulation and analysis of a diabetes dataset, are conducted in Section 4 to investigate finite-sample performance of the proposed approach. The article concludes with discussion in Section 5. Some technical details are provided in Appendix.

2. Iterated penalized variable selection

Without loss of generality, we assume that $E(Y) = 0$ and $E(X_j) = 0, j = 1, \dots, p_1$. In practice, we center $Y, X_1, \dots, X_{p_1}$ to achieve zero means. Under this condition and the zero expectations of f_j , $\alpha = 0$. To make covariates more comparable, in practice, we also rescale so that $\text{var}(X_j) = 1, j = 1, \dots, p_1$. For the nonparametric covariate effects $f_{p_1+1}, \dots, f_p$, we adopt a sieve approach (Chen 2007). Particularly, we assume that

$$f_j(X_j) = \sum_{k=1}^{s_j} \beta_j^k B_j^k(X_j).$$

Here s_j is the number of basis functions, $\beta_j = (\beta_j^1, \dots, \beta_j^{s_j})^T$ is the vector of unknown regression coefficients, and $B_j = (B_j^1, \dots, B_j^{s_j})^T$ is the set of known basis functions. In nonparametric estimation, s_j are allowed to diverge with sample size. With the basis expansion, model (1) can be written as

$$Y = \sum_{j=1}^p Z_j^T \beta_j + \varepsilon = Z^T \beta + \varepsilon,$$

where for $j = 1 \dots p_1$, $Z_j = X_j$ and for $j = p_1 + 1, \dots, p$, $Z_j = B_j$. $Z = (Z_1^T, \dots, Z_p^T)^T$. For $j = 1, \dots, p_1$, set $s_j = 1$. The dimensionality of the unknown parameter $\beta = (\beta_1^T, \dots, \beta_p^T)^T$ is $q = \sum_{j=1}^p s_j$. Assume that there are n i.i.d observations $\{(Y_i, Z_i); i = 1, \dots, n\}$. The OLS (ordinary least squares) objective function is

$$R(\beta) = \sum_{i=1}^n (Y_i - Z_i^T \beta)^2. \quad (2)$$

2.1. Iterated penalization

Consider the penalized estimate

$$\tilde{\beta} = \argmin_{\beta} \left\{ R(\beta) + \lambda_1 \sum_{j=1}^p \frac{1}{\sqrt{s_j}} \|\beta_j\|_2 \right\}. \quad (3)$$

Here λ_1 is the data-dependent tuning parameter. For $j = 1, \dots, p_1$, $\|\beta_j\|_2 = |\beta_j|$, and for $j = p_1 + 1, \dots, p$, $\|\beta_j\|_2 = (\sum_{k=1}^{s_j} (\beta_j^k)^2)^{1/2}$. That is, the proposed penalty is a mixture of the Lasso penalty for parametric covariate effects and the group Lasso penalty for nonparametric

covariate effects. We rescale the group L_2 norm by $\frac{1}{\sqrt{s_j}}$ so that the estimation is not dominated by large groups (covariate effects with a large number of basis functions). With parametric regression models, Zhao and Yu (2006), Zhang and Huang (2008) and others show that the Lasso penalization is not variable selection consistent unless irrepresentable conditions hold. More specifically, the Lasso tends to over-select. The intuition is that the Lasso employs too much penalty to large regression coefficients. A remedy is the adaptive Lasso approach proposed in Zou (2006), which adjusts the Lasso penalty using a set of data-dependent weights.

Motivated by the spirit of adaptive Lasso, we propose the iterated estimate

$$\widehat{\beta} = \argmin_{\beta} \left\{ R(\beta) + \lambda_2 \sum_{j=1}^p \frac{w_j}{\sqrt{s_j}} \|\beta_j\|_2 \right\}. \quad (4)$$

Here λ_2 is the data-dependent tuning parameter. $\{w_j; j = 1, \dots, p\}$ is a set of data-dependent weights which adjust the relative degree of penalization. A convenient choice for weights is

$w_j = \frac{1}{\|\tilde{\beta}_j\|_2^\gamma}$ where $\tilde{\beta}_j$ is the j th component of $\tilde{\beta}$. A wide range of γ values lead to similar asymptotic results. For simplicity, we set $\gamma = 1$.

The proposed iterated approach has been motivated by the following considerations. With parametric regression models, Zhang and Huang (2008) and others show that the Lasso penalized estimates have the following main properties. First, the number of nonzero estimates is bounded by $\min(n, p)$. With an iterated approach and the convention $\frac{1}{0} = \infty$, the second step of penalized estimation deals with a much smaller number of covariates when $n \ll p$. Second, it can select all of the true positives (covariates with nonzero effects) plus a few false positives; In addition, it is L_2 estimation consistent. In Section 3, we show that similar properties hold for the estimate defined in (3) for semiparametric regression models. With the over-selection and consistent-estimation properties, the Lasso penalized estimate makes an appropriate choice for initial estimate. Particularly, it satisfies all of the requirements for the reciprocal of weights in the adaptive Lasso approach (Zou 2006). The idea of achieving selection consistency via iterated penalization shares a similar spirit with

that in Meinshausen (2007), Meinshausen and Bühlmann (2006) and others. The present study advances from published studies by considering semiparametric regression models. In (4), $\lambda_2 = \lambda_1$ is sufficient to achieve asymptotic selection consistency. However, our numerical experience suggests that allowing different tuning parameters may improve finite-sample variable selection performance. The proposed approach only involves two Lasso (group Lasso)-type estimates, which can be realized efficiently using the algorithm described in Section 2.2.

2.2. Computational algorithm

In the proposed iterated procedure, the second step is a weighted version of the first step and can be computed using the same algorithm with simple rescaling. Below we describe the coordinate descent algorithm for computing the estimate defined in (3). This algorithm has been motivated by those in Meier et al. (2008) and Friedman et al. (2010).

Algorithm

-
- 1 Initialize $\beta = 0$ (component-wise);
 - 2 For $j = 1 \dots p$,
 - a. Denote β as the current estimate. Define a length $\sum_{j=1}^p s_j$ vector β^{-j} , where $\beta_i^{-j} = \beta_i$ if $i \neq j$ and $\beta_i^{-j} = 0$;
 - b. If $\left\| \sum_{i=1}^n Z_{i,j}(Y_i - Z_i^T \beta^{-j}) \right\|_2 \leq \frac{\lambda_1}{\sqrt{s_j}}$, update $\beta_j = 0$ (that is, all the s_j components are equal to 0). Here $Z_{i,j}$ is the j th component of Z on the i th subject;
 - c. Otherwise, define $\tau = \max(\min(\text{diag}(\sum_i Z_{i,j} Z_{i,j}^T)), C_*)$, where as suggested in Meier et al. (2008), $C_* > 0$ is small positive constant to ensure convergence. Update

$$\beta_j = \frac{\sum_i Z_{i,j} Y_i}{\left\| \sum_i Z_{i,j} Y_i \right\|_2} \frac{1}{\tau} \left(\left\| \sum_i Z_{i,j} Y_i \right\|_2 + \frac{\lambda_1}{\sqrt{s_j}} \right);$$
 - 3 Repeat Step 2 until convergence. In our numerical study, we use the L_2 norm of the difference between two consecutive estimates smaller than 0.001 as the criterion for convergence.
-

The coordinate descent algorithm computes penalized estimation with respect to one covariate at a time and cycles through all covariates. It transforms a complicated optimization problem to a series of simple ones. For optimization with each covariate, Step (2.c) shows a simple update formula. The main difference from the algorithm in Friedman et al. (2010) is that here the groups arise from basis expansions of the same covariates. Thus it is not feasible to transform and obtain orthonormal designs. We adopt the strategy proposed in Meier et al. (2008) and minimize over β_j along a special direction of the Hessian matrix in Step (2.c). In the penalized objective function, the first term has a simple least squares form and is continuously differentiable and regular in the sense of Tseng (2001). The proposed penalty is separable. Thus, the coordinate descent algorithm converges to a coordinate-wise minimum of the first term, which is also a stationary point (Tseng 2001). In our numerical study, convergence is achieved for all datasets within twenty iterations. The proposed algorithm only involves simple calculations and thus has a low computational cost.

2.3. Tuning parameter selection

The tuning parameters λ_1 and λ_2 jointly balance sparsity and goodness-of-fit. As $\lambda_1, \lambda_2 \rightarrow \infty$, fewer covariates are identified as associated with response. We propose selecting the tuning parameters using V-fold cross validation, which proceeds as follows. For a fixed λ_1 (λ_2), (1) Randomly partition data into V non-overlapping equal-sized subsets; (2) Remove

subset v ; Compute estimate $\hat{\beta}^{-v}$ using the remaining $V-1$ subsets; (3) Compute $R^v(\hat{\beta}^{-v})$, which is $R(\beta)$ evaluated on subset v using estimate $\hat{\beta}^{-v}$; (4) Repeat Steps (2)–(3) over all V subsets and compute the cross validation score as $CV(\lambda_1) = \sum_v R^v(\hat{\beta}^{-v})$. The optimal tuning is defined as

$$\lambda_{1,opt} = \argmin CV(\lambda_1).$$

In our numerical study, we set $V = 4$ and search over the discrete grid of $2 \dots -2, -1, 0, 1, 2 \dots$ for λ_1 and λ_2 .

In Section 3, we provide conditions on tuning under which the selection consistency holds. As a limitation of this study, we are unable to prove that the V -fold cross validated tunings satisfy such conditions. Tuning parameter selection with high-dimensional data is extremely challenging. Several recent studies investigate the optimality of BIC-type tuning parameter selection, however, under data and model settings much simpler than the present one. V -fold cross validation is used here because of its computational simplicity and satisfactory numerical performance. In V -fold cross validation, V is another “tuning parameter”. $V=4$ has been adopted in many published studies. When the sample size is very small, different choices of V may lead to different results. In our numerical study, with sample size at least 100, $V=4, \dots, 10$ lead to similar results. We refer to Chapter 7 of Hastie et al. (2009) for more detailed discussions on V -fold cross validation.

3. Asymptotic selection property

In this section, we investigate variable selection property of the proposed approach. We focus on the scenario where the dimensionality of Z is allowed to diverge with n .

With the convention $\frac{1}{0} = \infty$, variables not selected in the first step will not be selected in the second step. Thus, to investigate the final selection property, it is critical to first establish the selection property of the estimate defined in (3).

Denote $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. Let $\mathbf{Z} = (Z_1, \dots, Z_n)^T$ denote the $n \times q$ covariate matrix. Let $\mathbf{Z}_1, \dots, \mathbf{Z}_p$ be the p sub-matrices of $\mathbf{Z}$. Here the dimensionality of $\mathbf{Z}_j$ is $n \times s_j$. For an index set $A \subset \{1, \dots, p\}$, let $\mathbf{Z}_A = (\mathbf{Z}_j, j \in A)$ be the matrix with sub-matrices $\mathbf{Z}_j$ for $j \in A$. Denote $\sum_A = \mathbf{Z}_A^T \mathbf{Z}_A / n$. Define $|A| = \sum_{j=1}^p I(j \in A)$. Let $\beta^0 = (\beta_1^{0T}, \dots, \beta_p^{0T})^T$ be the unknown true value of β . Let $A^n = \{j: \|\beta_j^{0T}\|_2 \neq 0\}$ be the set of covariates with nonzero effects. Assume that

- (A1) There exists a constant K_1 such that $|A^n| < K_1 < \infty$;
- (A2) (a) The observations $\{(Y_i, Z_i), i = 1 \dots n\}$ are independent and identically distributed; (b) The random errors $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed with mean 0 and variance σ^2 . In addition, the random errors are subgaussian in the sense that there exist $K_2, K_3 > 0$ such that the tail probabilities of ε_i satisfy $P(|\varepsilon_i| > u) \leq K_2 \exp(-K_3 u^2)$ for all $u \geq 0$ and all i ;
- (A3) The covariates are bounded. That is, there exists a constant K_4 such that $|Z_{ij}| \leq K_4 < \infty$ for all i, j ;
- (A4) The sparse Riesz condition (SRC) is satisfied with rank q^* . That is, there exist constants $0 < c_l < c_u < \infty$ such that for $C = c_u/c_l$ and $q^* = (3 + 4C)|A^n|$, with probability converging to 1, $c_l \leq \mathbf{v}^T \sum_A \mathbf{v} \leq c_u$ for $\forall A$ with $|A| = q^*$, $\mathbf{v} \in \mathbb{R}^{\sum_{j=1}^p I(j \in A) s_j}$ and $\|\mathbf{v}\|_2 = 1$;

Similar assumptions have been made in other high-dimensional regression analysis studies. (A1) assumes that the model is sparse, in the sense that although the total number of covariates may diverge with sample size, the number of covariates with nonzero effects is small. Such an assumption is reasonable in practice where even though a large number of covariates may be measured, only a few are important. Assumption (A1) can be released in different ways. One possibility is to drop this condition. However the price is that $p \sim n^\eta$ with $\eta < 1$, a rate much lower than that obtained in Theorem 1. Another possibility is to assume *sparsifiable* (as opposed to sparse) model (Negahban et al. 2011). With sparsifiable models, only coefficients above a certain cutoff can be identified, which, loosely speaking, is equivalent to view the small nonzero coefficients as random noise. The subgaussian tail probability has been commonly assumed in regression analysis. It is weaker than the commonly assumed normal error. The boundedness assumption in (A3) is for theoretical convenience. The actual bound does not need to be known. In practical data analysis, the boundedness assumption can be easily satisfied. In theory, this assumption may limit the applicability of the proposed approach as it rules out some common distributions with infinite support. If this condition is released, other assumptions (for example bounded random errors) will have to be made. The SRC condition in (A4) is proposed in Zhang and Huang (2008) for Lasso with parametric regression models. This condition ensures that all eigenvalues of $\mathbf{Z}_B^T \mathbf{Z}_B / n$ with $|B| \leq q^*$ lie between c_l and c_u , and so the model corresponding to index set B is identifiable.

Consider $\tilde{\beta} = (\tilde{\beta}_1^T, \dots, \tilde{\beta}_p^T)^T$ defined in (3). Define $\tilde{A} = \{j: \|\tilde{\beta}_j\|_2 > 0\}$. The main selection results are summarized as follows.

Theorem 1—Suppose that assumptions (A1)–(A4) are satisfied. In addition,

$\lambda_1 \geq O(1) \sqrt{n \log(q)}$. Then

- i. $P(|\tilde{A}| \leq (2 + 4C)|A^n|) \rightarrow 1$;
- ii. If $\lambda_1/n \rightarrow 0$ and $\log(q)/n \rightarrow 0$, then with probability converging to 1, all covariates with nonzero effects are selected;
- iii. $\|\tilde{\beta} - \beta^0\|_N^2 \leq \frac{4\lambda_1^2 |A^n|}{n^2 c_l^2} + O_p\left(\frac{|\tilde{A}| \log(q)}{nc_l^2}\right)$. Particularly, if $\lambda_1 = O(\sqrt{n \log(q)})$, then $\|\tilde{\beta} - \beta^0\|_N^2 = O_p(\log(q)/n)$. Here $\|\cdot\|_N$ is the normalized L_2 norm defined in Appendix.

Theorem 1 suggests that with probability converging to 1, the number of covariates selected in the initial estimate (3) is a finite multiply of the number of covariates truly with nonzero effects. Under assumption (A1), even as $n, q \rightarrow \infty$, the number of selected covariates is asymptotically bounded. Theorem 1 also suggests that all covariates with nonzero effects are selected with probability converging to 1, and the estimate is L_2 consistent, making it an appropriate choice for weights in the iterated estimate (4).

We now consider the iterated estimate. For a vector $a = (a_1, \dots, a_p)$, define its sign vector as $\text{sgn}(a) = (\mathbb{I}(\|a_1\|_2 > 0), \dots, \mathbb{I}(\|a_p\|_2 > 0))$.

Theorem 2—Suppose that assumptions (A1)–(A4) are satisfied. In addition, $\log(q)/n \rightarrow 0$ and $\lambda_2 = O(\sqrt{n \log(q)})$. Then

$$P(\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^0)) \rightarrow 1.$$

Theorem 2 shows that the iterated estimate is selection consistent in the sense that it can correctly distinguish covariates with zero effects from those with nonzero effects. For the validity of Theorems 1 and 2, conditions on q , the total number of unknown parameters, need to be assumed. Consider the simplified scenario where there is no parametric covariate effect. Then such conditions state that there needs to be a “balance” between the number of basis functions per nonparametric covariate effect and the number of nonparametric covariate effects. Such a constraint is intuitively reasonable and has been assumed in nonparametric studies with comparable settings.

4. Numerical study

4.1. Simulation

We conduct simulation to investigate finite sample performance of the proposed approach. As shown in Table 1, we simulate 100 or 200 iid observations. We consider two scenarios with $p = 50$ or 500 covariates. We first simulate the covariates to have joint normal distributions with marginal mean equal to zero and standard deviation equal to 0.5. Covariates i and j have correlation coefficient $\rho^{|i-j|}$, that is, the auto-regressive correlation structure. We consider two ρ values, representing weak and strong correlations. We then add floor -3 and ceiling 3 to the covariates to satisfy the boundedness condition. When $p = 50$, we assume that there are ten nonparametric covariate effects and forty parametric covariate effects. Among the forty parametric covariate effects, ten are nonzero, all with regression coefficient equal to 2. Among the ten nonparametric covariate effects, there are five nonzero ones. Particularly, there are three copies of function $f_1(X) = \sin(\frac{\pi}{2}X - \frac{\pi}{2})$ and two copies of function $f_2(X) = 5(\frac{x}{8} + 0.1)^2 + 5.7\exp(-\frac{x}{8} - 0.5)$. We note that in execution, those two functions need to be normalized to have mean zero. When $p = 500$, we assume that there are twenty nonparametric covariate effects and 480 parametric covariate effects. Among the 480 parametric covariate effects, twenty are nonzero, all with regression coefficient equal to 2. Among the twenty nonparametric covariate effects, there are ten nonzero ones. Particularly, there are six copies of function $f_1(X)$ and four copies of function $f_2(X)$. The random errors are generated from a normal distribution with variance equal to 2 or 4.

When implementing the proposed approach, we choose the B-spline basis functions for the estimation of smooth nonparametric covariate effects (Schumaker 1981). We adopt the approach in Xiang and Wahba (1997), which states that an estimate with the number of basis functions growing at the rate of $n^{1/5}$ can achieve the same asymptotic precision as the full space. In practical data analysis, we suggest at least five basis functions. To better gauge performance of the proposed approach, we also consider the following three alternatives:

M₂: which assumes that all covariate effects are parametric. An iterated Lasso penalization approach similar to the proposed one is adopted for estimation and variable selection;

M₃: which is a full oracle approach. Here, we know exactly which covariate effects are nonzero. In addition, we know which nonzero covariate effects are nonparametric and which are parametric. As the total number of unknown parameters is smaller than sample size, a simple ordinary least squares (OLS) estimation is conducted;

M₄: which is a partial oracle approach. Here, we know exactly which covariate effects are nonzero. However, we assume that all covariate effects are parametric and adopt a simple OLS estimation.

Among the three approaches, **M₂** is a modification of the commonly adopted Lasso approach and serves as a benchmark to the proposed approach. It is designed to demonstrate the importance of semiparametric modeling and estimation. **M₃** and **M₄** are oracle

approaches, use information we do not have in practice, and are only feasible in simulation. In the comparison of different approaches, we evaluate identification as well as prediction performance. In the evaluation of identification accuracy, we are interested in the number of covariate effects identified and number of true positives. With semipara-metric models, we evaluate the parametric and nonparametric components separately. In the evaluation of prediction performance, we generate independent data under the same settings and compute the PMSE (prediction mean squared error). Another possible performance measure is the estimation accuracy of regression coefficients. In practical data analysis, penalized estimates demonstrate shrinkage towards zero. The PMSE can reflect the tradeoff between bias and variance and be a better measure than estimation accuracy.

Simulation results based on 200 replicates are shown in Table 1 where we refer to the proposed approach as $\mathbf{M}_1$. Under all simulation scenarios, the proposed approach outperforms the benchmark $\mathbf{M}_2$. Particularly, both approaches are capable of identifying the true nonzero parametric covariate effects, with the proposed approach identifying fewer false positives. The proposed approach significantly outperforms $\mathbf{M}_2$ by identifying more nonzero nonparametric covariate effects. We note that although the proposed approach has the asymptotic consistency property, in practice, there may still be a number of false positives and false negatives. Similar phenomena have been observed with other penalization approaches. In terms of prediction, the proposed approach has PMSEs much smaller than $\mathbf{M}_2$. We note that the PMSEs of $\mathbf{M}_3$, the full oracle approach, are larger than those of the proposed approach and $\mathbf{M}_4$. This is caused by a large number of unknown parameters from basis expansions, small sample size and possible colinearity. We also briefly look into estimation property of the nonparametric covariate effects. In Figure 1, for the simulation scenario described in row 13 of Table 1, we show the mean estimates and point-wise 95% confidence intervals for estimation of the nonparametric covariate effects. Because of penalization, we clearly see shrinkage of the estimates. On average, the proposed approach provides satisfactory estimation of the nonparametric covariate effects.

4.2. Analysis of diabetes data

The diabetes data is courtesy of Dr. John Schorling, Department of Medicine, University of Virginia School of Medicine (Willems et al. 1997; Schorling et al. 1997). The data consists of 19 variables on 403 subjects from 1,046 African Americans who were interviewed in a study to investigate 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 who were screened for diabetes. Glycosolated hemoglobin > 7.0 is usually taken as a positive diagnosis of this disease.

We consider Glycosolated hemoglobin as the response variable and the other 16 variables as covariates. Brief descriptions of the covariates are provided in Table 2 and Figure 2. Among the covariates, three are categorical (location, gender, frame) and naturally correspond to parametric covariate effects. For the continuously distributed covariates, our preliminary analysis suggests possible nonlinear effects for *stab.glu*, *chol*, *hdl* and *ratio*. In our analysis, we remove subjects with missing measurements, leading to an effective sample size of 366. We employ the proposed approach and identify four nonzero parametric covariate effects: *age*, *waist*, *time.ppn* and *frame.m* (Table 2), all of which are positively associated with the response variable. This finding is consistent with those in the literature. The estimates of nonparametric covariate effects are shown in Figure 2. Three out of the four covariates have nonzero nonparametric covariate effects. All of the three nonzero nonparametric covariate effects seriously deviate from linear and do not seem to have simple parametric approximations.

As methods M_3 and M_4 are not feasible with practical data, we analyze using method M_2 . In Table 2, we can see that M_2 identifies two more parametric covariate effects (bp.l.s and loca). M_2 also suggests zero effect for HDL. For stab.glu, chol and ratio, the M_2 estimates are all positive, suggesting similar overall trends as those shown in Figure 2.

With real data, it is impossible to objectively evaluate variable selection accuracy. We conduct evaluation of prediction performance, which can provide indirect evaluation of variable selection results. Particularly, if the variables selected are meaningful, prediction using these variables should be accurate. The prediction evaluation proceeds as follows: (1) Random split data into a training set with size 200 and a testing set with size 166; (2) Conduct penalized estimation, variable selection and model construction using the training set; (3) Make prediction for subjects in the testing set using the training set estimate. Compute PMSE; (4) Repeat Steps (1)–(3) 200 times and compute the mean of PMSEs. With the proposed approach, the mean of PMSEs is 0.467 with standard deviation 0.063. As a comparison, we consider two alternative approaches. One is M_2 , and the other is a ridge approach with tuning selected using V-fold cross validation. With M_2 , the mean of PMSEs is 0.475 with standard deviation 0.092; With ridge, the mean of PMSEs is 0.495 with standard deviation 0.074. The proposed approach outperforms the two alternatives with a smaller prediction error.

5. Discussion

Semiparametric regression with multiple covariates is commonly encountered. The estimation aspect of such models has been investigated in a large number of publications, whereas not enough attention has been paid to the variable selection aspect. In this article, we propose an iterated penalization approach, which is built on a mixture of the Lasso penalty for parametric covariate effects and the group Lasso penalty for nonparametric covariate effects. We show that the proposed approach has the desired variable selection consistency property, even when the total number of unknown parameters diverges with sample size. With the selection consistency and under assumption (A1), estimation properties follow from standard semiparametric estimation results. Simulation study and data analysis show satisfactory finite-sample performance of the proposed approach.

As discussed in the Introduction section, there are several other approaches that can be used for a similar purpose. The proposed approach can be preferred because of its lucid asymptotic properties, computational simplicity and satisfactory numerical performance. Comprehensive numerical comparisons of different approaches are beyond the scope of this article, especially considering that some of the alternative approaches do not have publicly available software. The proposed penalty is built on the Lasso penalty. We conjecture that it is possible to follow a similar strategy and build penalties based on, for example, bridge and SCAD. The Lasso is selected mostly because of its computational simplicity. In theory, the proposed approach can accommodate multiple different forms of basis functions. In numerical study, we assume smooth covariate effects, as most biological processes are continuous. B-spline basis functions are chosen because of their popularity and satisfactory numerical performance. Using other basis functions should be straightforward. With the proposed approach, we select tunings using V-fold cross validation. Numerically, we have experimented with a few other tuning parameter selection methods including AIC, BIC and GCV, and find that none significantly outperforms V-fold cross validation. The theoretical optimality of tuning parameter selection is extremely challenging and deferred to future research. In this study, with high dimensional data, we focus on the variable selection aspect. A relevant aspect not studied is inference. Inference in penalized estimation is a hard problem. We have searched literature and found that, in theory, there are some distributional results. However, numerical studies have shown conflicting performance of those

approaches. There are several recent studies suggesting ad hoc ways of inference. However, there is a lack of consensus. Sixteen different simulation scenarios are presented in Table 1. We have experimented with a few other simulation settings and reached similar conclusions (results omitted).

Acknowledgments

The authors would like to thank the associate editor and two referees for careful review and insightful comments, which have led to a significant improvement of the manuscript. This study has been supported by awards DMS-0904181 from NSF and CA142774 from NIH, USA.

References

- Chen X. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*. 2007; Chapter 76
- Du P, Ma S, Liang H. Penalized variable selection procedure for Cox models with semiparametric relative risk. *Annals of Statistics*. 2010; 38(4):2092–2117. [PubMed: 20802853]
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *JASA*. 2001; 96:1348–1360.
- Friedman, J.; Hastie, T.; Tibshirani, R. Technical Report. Department of Statistics, Stanford University; 2010. A note on the group lasso and a sparse group lasso.
- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer; 2009.
- Horowitz, J. *Semiparametric and Nonparametric Methods in Economics*. New York: Springer; 2009.
- Huang J, Ma S, Xie H, Zhang C. A group bridge approach for variable selection. *Biometrika*. 2009; 96(2):339–355. [PubMed: 20037673]
- Huang J, Ma S, Zhang C. Adaptive Lasso for high dimensional regression models. *Statistica Sinica*. 2008; 18:1603–1618.
- Huang J, Horowitz J, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models'. *Annals of Statistics*. 2008; 36:587–613.
- Li R, Liang H. Variable selection in semiparametric regression modeling. *Annals of Statistics*. 2008; 36:261–286. [PubMed: 19122808]
- Meier L, van de Geer S, Bühlmann P. The group lasso for logistic regression. *JRSSB*. 2008; 70:53–71.
- Meinshausen N. Relaxed Lasso. *Computational Statistics and Data Analysis*. 2007; 52:374–393.
- Meinshausen N, Bühlmann P. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*. 2006; 34:1436–1462.
- Negahban, S.; Ravikumar, P.; Wainwright, MJ.; Yu, B. Manuscript, University of California; Berkeley: 2011. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers.
- Ruppert, D.; Wand, MP.; Carroll, RJ. *Semiparametric Regression*. Cambridge: Cambridge University Press; 2003.
- Schorling JB, Roach J, Siegel M, Baturka N, Hunt DE, Guterbock TM, Stewart HL. A trial of church-based smoking cessation interventions for rural African Americans. *Preventive Medicine*. 1997; 26:92–101. [PubMed: 9010903]
- Schumaker, L. *Spline Functions: Basic Theory*. New York: Wiley; 1981.
- Tibshirani R. Regression shrinkage and selection via the lasso. *JRSSB*. 1996; 58(1):267–288.
- Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *J Optimiz Theory App*. 2001; 109:475–494.
- Van de Geer S. High dimensional generalized linear models and the Lasso. *Annals of Statistics*. 2008; 36:614–645.
- Van der Vaart, AW.; Wellner, JA. *Weak Convergence and Empirical Processes*. New York: Springer Verlag; 1996.
- Wand MP. Smoothing and mixed models. *Computational Statistics*. 2003; 18:223–249.

- Willems JP, Saunders JT, Hunt DE, Schorling JB. Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. *Southern Medical Journal*. 1997; 90:814–820. [PubMed: 9258308]
- Xiang, D.; Wahba, G. Approximate smoothing spline methods for large data set in the binary case. *Proceedings of the 1997 ASA Joint Statistical Meetings, Biometrics Section*; 1997. p. 94–98.
- Xie H, Huang J. SCAD-penalized regression in high-dimensional partially linear models. *Annals of Statistics*. 2009; 37:673–696.
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *JRSSB*. 2006; 68:49–67.
- Zhang C. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*. 2010; 38:894–942.
- Zhang C, Huang J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*. 2008; 36:1567–1594.
- Zhao P, Yu B. On model selection consistency of Lasso. *Journal of Machine Learning Research*. 2006; 7:2541–2563.
- Zou H. The adaptive Lasso and its oracle properties. *JASA*. 2006; 102:1418–1429.

Appendix

Consider $\varepsilon_i = Y_i - Z_i^T \beta^0$. Define $\xi_j = \sum_{i=1}^n Z_{i,j} \varepsilon_i$, $1 \leq j \leq q$. Let $\xi_n = \max_{1 \leq j \leq q} |\xi_j|$.

Lemma 1

Suppose that assumptions (A2) and (A3) are satisfied. Then

$$E(\xi_n) \leq C_1 \sqrt{\log(q)} \left(\sqrt{2C_2 n \log(q)} + 4\log(2q) + C_2 n \right)^{1/2},$$

where C_1 and C_2 are two positive constants. In addition, if $\log(q)/n \rightarrow 0$,

$$E(\xi_n) = O(1) \sqrt{n \log q}.$$

Proof

Let $s_{nj}^2 = \sum_{i=1}^n Z_{i,j}^2$. Conditional on $Z_{i,j}$'s, assumptions (A2) and (A3) imply that ξ_j 's are subgaussian. Let $s_n^2 = \max_{1 \leq j \leq q} s_{nj}^2$. By (A2) and the maximal inequality for subgaussian random variables (Van der Vaart and Wellner 1996, Lemmas 2.2.1 and 2.2.2),

$$E \left(\max_{1 \leq j \leq q} |\xi_j| \mid Z_{i,j}, 1 \leq i \leq n, 1 \leq j \leq q \right) \leq C_1 s_n \sqrt{\log(q)},$$

for a constant $C_1 > 0$. Therefore,

$$E \left(\max_{1 \leq j \leq q} |\xi_j| \right) \leq C_1 \sqrt{\log(q)} E(s_n). \quad (\text{A1})$$

Since

$$\sum_{i=1}^n E(Z_{i,j}^2 - EZ_{i,j}^2)^2 \leq 4C_2n, \quad (\text{A2})$$

and

$$\max_{1 \leq j \leq q} \sum_{i=1}^n EZ_{i,j}^2 \leq C_2n, \quad (\text{A3})$$

by Lemma 4.2 of Van de Geer (2008), (A2) implies

$$E \left(\max_{1 \leq j \leq q} \left| \sum_{i=1}^n \{Z_{i,j}^2 - EZ_{i,j}^2\} \right| \right) \leq \sqrt{2C_2n \log(q)} + 4\log(2q).$$

Therefore, by (A3) and the triangle inequality,

$$Es_n^2 \leq \sqrt{2C_2n \log(q)} + 4\log(2q) + C_2n.$$

Now since $Es_n \leq (Es_n^2)^{1/2}$, we have

$$Es_n \leq \left(\sqrt{2C_2n \log(q)} + 4\log(2q) + C_2n \right)^{1/2}. \quad (\text{A4})$$

The lemma follows from (A1) and (A4).

Proof of Theorem 1

Part (i) follows from the proof of Theorem 1 of Zhang and Huang (2008). The difference is that here we use the subgaussian assumption to control the tail probabilities, as opposed to the normality condition assumed in Zhang and Huang (2008). Since subgaussian random variables have the same tail behavior as normal random variables, the argument of Zhang and Huang (2008) goes through.

Part (ii) follows from part (iii) and the assumption that the number of nonzero covariate effects is fixed. Thus the L_2 norms of the nonzero covariate effects are bounded away from 0 by a positive constant independent of n .

We now prove part (iii) of Theorem 1. Denote $\|\beta_j\|_N = \frac{1}{\sqrt{c_j}} \|\beta_j\|_2$ as the normalized L_2 norm. From the definition of $\tilde{\beta}$,

$$\|Y - Z\tilde{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^p \|\tilde{\beta}_j\|_N \leq \|Y - Z\beta^0\|_2^2 + \lambda_1 \sum_{j=1}^p \|\beta_j^0\|_N.$$

That is,

$$\|Y - Z\tilde{\beta}\|_2^2 + \lambda_1 \sum_{j \in A^n} \|\tilde{\beta}_j\|_N \leq \|Y - Z\beta^0\|_2^2 + \lambda_1 \sum_{j \in A^n} \|\beta_j^0\|_N.$$

This result leads to

$$\|Y - Z\tilde{\beta}\|_2^2 - \|Y - Z\beta^0\|_2^2 \leq \lambda_1 \sum_{j \in A^n} \|\tilde{\beta}_j - \beta_j^0\|_N. \quad (\text{A5})$$

Denote $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$. The above result leads to

$$\|Z(\tilde{\beta} - \beta^0)\|_2^2 - 2\varepsilon^T Z(\tilde{\beta} - \beta^0) \leq \lambda_1 \sum_{j \in A^n} \|\tilde{\beta}_j - \beta_j^0\|_N.$$

Denote $B = A^n \cup \tilde{A} = \{j: \|\beta_j^0\|_2 > 0\} \cup \{j: \|\tilde{\beta}_j\|_2 > 0\}$. From part (i) of Theorem 1, $P(B|q^*) \rightarrow 1$, with q^* defined in assumption (A4). Denote $Z_B = (Z_j, j \in B)$, $\tilde{\beta}_B = (\tilde{\beta}_j, j \in B)$, and $\beta_B^0 = (\beta_j^0, j \in B)$. Denote

$$\eta_B = Z_B(\tilde{\beta}_B - \beta_B^0).$$

Since $A^n \subset B$,

$$\sum_{j \in A^n} \|\tilde{\beta}_j - \beta_j^0\|_N \leq \sqrt{|A^n|} \|\tilde{\beta}_{A^n} - \beta_{A^n}^0\|_N \leq \sqrt{|A^n|} \|\tilde{\beta}_B - \beta_B^0\|_N. \quad (\text{A6})$$

Combining (A5) and (A6),

$$\|\eta_B\|_2^2 - 2\varepsilon^T \eta_B \leq \lambda_1 \sqrt{|A^n|} \|\tilde{\beta}_B - \beta_B^0\|_N. \quad (\text{A7})$$

Define $\varepsilon_B^* = Z_B(Z_B^T Z_B)^{-1} Z_B^T \varepsilon$. We have

$$\varepsilon^T \eta_B = \varepsilon^T Z_B(\tilde{\beta}_B - \beta_B^0) = \{ (Z_B^T Z_B)^{-1/2} Z_B^T \varepsilon \}^T \{ (Z_B^T Z_B)^{1/2} (\tilde{\beta}_B - \beta_B^0) \}.$$

With the Cauchy-Schwarz inequality,

$$2|\varepsilon^T \eta_B| \leq 2\|\varepsilon_B^*\|_2 \cdot \|\eta_B\|_2 \leq 2\|\varepsilon_B^*\|_2^2 + \frac{1}{2}\|\eta_B\|_2^2. \quad (\text{A8})$$

Combining (A7) and (A8),

$$\|\eta_B\|_2^2 \leq 4\|\varepsilon_B^*\|_2^2 + 2\lambda_1 \sqrt{|A^n|} \|\tilde{\beta}_B - \beta_B^0\|_N \quad (\text{A9})$$

By the SRC condition (A4), $\|\eta_B\|_2^2 \geq nc_l \|\tilde{\beta}_B - \beta_B^0\|_N^2$. Thus (A9) implies

$$nc_l \|\tilde{\beta}_B - \beta_B^0\|_N^2 \leq 4\|\varepsilon_B^*\|_2^2 + \frac{(2\lambda_1 \sqrt{|A^n|})^2}{2nc_l} + \frac{1}{2}nc_l \|\tilde{\beta}_B - \beta_B^0\|_N^2.$$

It follows that

$$\|\tilde{\beta}_B - \beta_B^0\|_N^2 \leq \frac{8\|\varepsilon_B^*\|_2^2}{nc_l} + \frac{4\lambda_1^2 |A^n|}{n^2 c_l^2}. \quad (\text{A10})$$

We also have

$$\|\varepsilon_B^*\|_2^2 = \|(\mathbf{Z}_B^T \mathbf{Z}_B)^{-1/2} \mathbf{Z}_B^T \varepsilon\|_2^2 \leq \frac{1}{nc_l} \|\mathbf{Z}_B \varepsilon\|_2^2 \leq \frac{1}{nc_l} \max_{A: |A| \leq q^*} \|\mathbf{Z}_A \varepsilon\|_2^2.$$

We have

$$\max_{A: |A| \leq q^*} \|\mathbf{Z}_A^T \varepsilon\|_2^2 = \max_{A: |A| \leq q^*} \sum_{j \in A} |\mathbf{Z}_j^T \varepsilon|^2 \leq q^* \max_{1 \leq j \leq q} |\mathbf{Z}_j^T \varepsilon|^2.$$

By Lemma 1,

$$\max_{1 \leq j \leq q} |\mathbf{Z}_j^T \varepsilon|^2 = n \max_{1 \leq j \leq q} |n^{-1/2} \mathbf{Z}_j^T \varepsilon|^2 = O_p(n \log(q)).$$

Therefore,

$$\|\varepsilon_B^*\|_2^2 = O_p\left(\frac{q^* \log q}{c_l}\right). \quad (\text{A11})$$

The result follows from (A10) and (A11).

Proof of Theorem 2

The proof mainly follows from Huang et al. (2008). We briefly outline the proof below. By the Karush-Kuhn-Tucker conditions, $\widehat{\beta} = (\widehat{\beta}_1^T, \dots, \widehat{\beta}_p^T)^T$ is the unique solution of the estimate defined in (4) if

$$\begin{cases} \mathbf{Z}_j^T(\mathbf{Y}-\mathbf{Z}\widehat{\boldsymbol{\beta}})=\lambda_2 \frac{1}{\sqrt{s_j}\|\widehat{\boldsymbol{\beta}}_j\|_2} \text{sgn}(\widehat{\boldsymbol{\beta}}_j), & \|\widehat{\boldsymbol{\beta}}_j\|_2 > 0 \\ \|\mathbf{Z}_j^T(\mathbf{Y}-\mathbf{Z}\widehat{\boldsymbol{\beta}})\|_2 \leq \lambda_2 \frac{1}{\sqrt{s_j}\|\widehat{\boldsymbol{\beta}}_j\|_2}, & \|\widehat{\boldsymbol{\beta}}_j\|_2 = 0 \end{cases} \quad (\text{A12})$$

and the vectors $\{\mathbf{Z}_j, \|\widehat{\boldsymbol{\beta}}_j\|_2 > 0\}$ are linearly independent. With slight abuse of notation, here

$\text{sgn}(\widehat{\boldsymbol{\beta}}_j)$ denotes the component-wise sign function. Recall that $A^n = \{j: \|\boldsymbol{\beta}_j^0\|_2 > 0\}$. Let

$$\tilde{s}_1 = \left(\frac{1}{\sqrt{s_j}\|\widehat{\boldsymbol{\beta}}_j\|_2} \text{sgn}(\boldsymbol{\beta}_j^0), j \in A^n \right)^T \text{ and } \mathbf{Z}_{A^n} = (\mathbf{Z}_j, j \in A^n), \boldsymbol{\beta}_{A^n}^0 = (\boldsymbol{\beta}_j^0, j \in A^n)^T. \text{ Define}$$

$$\widehat{\boldsymbol{\beta}}_{A^n} = (\mathbf{Z}_{A^n}^T \mathbf{Z}_{A^n})^{-1} (\mathbf{Z}_{A^n}^T \mathbf{Y} - \lambda_2 \tilde{s}_1) = \boldsymbol{\beta}_{A^n}^0 + C_{11}^{-1} (\mathbf{Z}_{A^n}^T \boldsymbol{\varepsilon} - \lambda_2 \tilde{s}_1) / n \quad (\text{A13})$$

where $C_{11} = \mathbf{Z}_{A^n}^T \mathbf{Z}_{A^n} / n$. If $\text{sgn}(\widehat{\boldsymbol{\beta}}_{A^n}) = \text{sgn}(\boldsymbol{\beta}_{A^n}^0)$, then the equation in (A12) holds for $(\widehat{\boldsymbol{\beta}}_{A^n}^T, \mathbf{0}^T)^T$. Thus, since $\mathbf{Z}\widehat{\boldsymbol{\beta}} = \mathbf{Z}_{A^n} \widehat{\boldsymbol{\beta}}_{A^n}$ for this $\widehat{\boldsymbol{\beta}}$,

$$\text{sgn}(\widehat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta}^0) \quad \text{if} \quad \begin{cases} \text{sgn}(\widehat{\boldsymbol{\beta}}_{A^n}) = \text{sgn}(\boldsymbol{\beta}_{A^n}^0) \\ \|\mathbf{Z}_j^T(\mathbf{Y} - \mathbf{Z}_{A^n} \widehat{\boldsymbol{\beta}}_{A^n})\|_2 \leq \lambda_2 \frac{1}{\sqrt{s_j}\|\widehat{\boldsymbol{\beta}}_j\|_2} \forall j \notin A^n. \end{cases} \quad (\text{A14})$$

Let $H_n = \mathbf{I}_n - \mathbf{Z}_{A^n} C_{11}^{-1} \mathbf{Z}_{A^n}^T / n$. It follows from (A13) that

$$\mathbf{Y} - \mathbf{Z}_{A^n} \widehat{\boldsymbol{\beta}}_{A^n} = \boldsymbol{\varepsilon} - \mathbf{Z}_{A^n} (\widehat{\boldsymbol{\beta}}_{A^n} - \boldsymbol{\beta}_{A^n}^0) = H_n \boldsymbol{\varepsilon} + \mathbf{Z}_{A^n} C_{11}^{-1} \tilde{s}_1 \lambda_2 / n,$$

so that by (A14),

$$\text{sgn}(\widehat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta}^0) \quad \text{if} \quad \begin{cases} \|\boldsymbol{\beta}_j^0 - \widehat{\boldsymbol{\beta}}_j\|_2 \leq \|\boldsymbol{\beta}_j^0\|_2, & \forall j \in A^n \\ |\mathbf{Z}_j^T (H_n \boldsymbol{\varepsilon} + \mathbf{Z}_{A^n} C_{11}^{-1} \tilde{s}_1 \lambda_2 / n)| < \lambda_2 \frac{1}{\sqrt{s_j}\|\widehat{\boldsymbol{\beta}}_j\|_2}, & \forall j \notin A^n. \end{cases} \quad (\text{A15})$$

Thus, by (A15) and (A13),

$$\begin{aligned} P\{\text{sgn}(\widehat{\boldsymbol{\beta}}) \neq \text{sgn}(\boldsymbol{\beta}^0)\} &\leq P\{\|\mathbf{e}_j^T C_{11}^{-1} \mathbf{Z}_{A^n}^T \boldsymbol{\varepsilon}\|_2 / n \geq \|\boldsymbol{\beta}_j^0\|_2 / 2 \text{ for some } j \in A^n\} \\ &\quad + P\{\|\mathbf{e}_j^T C_{11}^{-1} \tilde{s}_1\|_2 \lambda_2 / n \geq \|\boldsymbol{\beta}_j^0\|_2 / 2 \text{ for some } j \in A^n\} \\ &\quad + P\{\|\mathbf{Z}_j^T H_n \boldsymbol{\varepsilon}\|_2 \geq \frac{2\lambda_2}{\sqrt{s_j}\|\widehat{\boldsymbol{\beta}}_j\|_2} \text{ for some } j \notin A^n\} \\ &\quad + P\{\|\mathbf{Z}_j^T \mathbf{Z}_{A^n} C_{11}^{-1} \tilde{s}_1\|_2 / n \geq \frac{2}{\sqrt{s_j}\|\widehat{\boldsymbol{\beta}}_j\|_2} \text{ for some } j \notin A^n\} \\ &= P\{B_{n1}\} + P\{B_{n2}\} + P\{B_{n3}\} + P\{B_{n4}\}, \end{aligned}$$

where $\mathbf{e}_j$ is the unit vector in the direction of the j -th covariate. Therefore, to prove the theorem, it suffices to show that each probability in the last line converges to zero. The same argument as in Huang et al. (2008) can be used here and is omitted.

Figure 1. Simulation study under the scenario corresponding to row 13 of Table 1: estimation of nonzero nonparametric covariate effects. True values (solid lines), median estimates (dash-dotted lines) and point-wise 95% confidence intervals (dotted lines) over 200 replicates.

Figure 2. Analysis of diabetes data: estimation of nonparametric covariate effects. stab.glu: stabilized glucose; chol: total cholesterol; hdl: High density lipoprotein; ratio: Cholesterol/HDL ratio.

Table 1

Simulation study: median (standard deviation) computed based on 200 replicates. Under $\mathbf{M}_1$ and $\mathbf{M}_2$, the identification results include number of parametric covariate effects identified, number of true parametric covariate effects identified, number of nonparametric covariate effects identified, number of true nonparametric covariate effects identified.

n	p	ρ	var(e)	M ₁		M ₂		PMSE	identification	PMSE	PMSE	PMSE	M ₃	PMSE	M ₄	PMSE
				identification		identification										
100	50	0.3	2	13 (1.2)	10 (0.0)	4 (1.0)	4 (0.9)	3.224 (0.644)	19 (4.1)	10 (0.0)	4 (1.4)	3 (1.0)	6.245 (0.798)	3.953 (1.197)	3.227 (0.407)	
			4	14 (2.1)	10 (0.0)	4 (1.0)	4 (1.0)	3.266 (0.834)	19 (4.5)	10 (0.0)	4 (1.9)	2 (1.0)	6.008 (0.906)	3.822 (1.411)	3.364 (0.703)	
	0.7	2	12 (1.3)	10 (0.0)	4 (1.2)	4 (0.9)	3.700 (0.638)	15 (3.2)	10 (0.0)	3 (1.5)	2 (1.1)	6.525 (0.966)	4.277 (1.158)	3.863 (0.706)		
		4	12 (1.4)	10 (0.0)	5 (1.2)	4 (0.8)	3.817 (0.666)	14 (3.1)	10 (0.0)	3 (1.3)	2 (1.0)	6.063 (1.103)	4.549 (1.180)	4.263 (0.650)		
500	0.3	2	42 (4.9)	20 (0.1)	5 (1.0)	5 (0.9)	13.829 (4.857)	57 (14.9)	20 (0.1)	3 (1.6)	2 (1.2)	25.977 (4.488)	11.214 (4.835)	6.634 (0.999)		
		4	55 (3.8)	20 (0.2)	3 (0.8)	3 (0.7)	16.819 (5.937)	70 (13.3)	20 (0.2)	3 (1.5)	2 (1.1)	25.838 (6.151)	10.632 (5.873)	6.955 (1.245)		
	0.7	2	43 (3.6)	20 (0.2)	5 (1.6)	5 (1.5)	11.163 (2.657)	46 (12.4)	20 (0.2)	3 (1.3)	2 (1.1)	22.518 (3.232)	11.828 (3.492)	10.197 (2.189)		
		4	45 (3.9)	20 (0.2)	3 (1.5)	3 (1.5)	12.352 (3.351)	43 (11.8)	20 (0.2)	3 (1.6)	2 (1.1)	21.284 (3.823)	12.571 (4.735)	9.810 (1.913)		
200	50	0.3	2	12 (1.4)	10 (0.0)	6 (0.5)	6 (0.5)	3.052 (0.312)	19 (3.8)	10 (0.0)	5 (1.7)	3 (1.0)	5.616 (0.669)	3.929 (1.062)	2.921 (0.416)	
			4	14 (1.6)	10 (0.0)	5 (0.7)	5 (0.6)	2.062 (0.357)	18 (4.8)	10 (0.0)	5 (1.8)	3 (0.9)	6.075 (0.829)	3.567 (0.996)	3.043 (0.404)	
	0.7	2	11 (1.1)	10 (0.0)	6 (0.6)	6 (0.4)	2.791 (0.302)	14 (3.7)	10 (0.0)	3 (1.2)	2 (0.8)	6.026 (0.651)	3.983 (0.864)	3.102 (0.392)		
		4	12 (1.7)	10 (0.0)	5 (0.8)	5 (0.5)	1.849 (0.351)	15 (3.7)	10 (0.0)	4 (1.4)	2 (0.9)	6.846 (0.814)	4.805 (1.496)	3.744 (0.471)		
500	0.3	2	27 (4.0)	20 (0.0)	7 (1.5)	7 (1.4)	7.180 (1.416)	50 (18.1)	20 (0.0)	5 (1.5)	4 (1.0)	17.763 (2.044)	10.456 (2.336)	4.953 (0.604)		
		4	40 (4.8)	20 (0.0)	3 (1.5)	3 (1.4)	7.467 (1.209)	64 (17.3)	20 (0.0)	5 (1.5)	4 (1.4)	18.571 (2.161)	11.104 (4.682)	5.427 (0.629)		
	0.7	2	26 (2.8)	20 (0.0)	8 (1.3)	8 (1.3)	6.964 (0.827)	35 (11.5)	20 (0.0)	4 (1.7)	3 (1.2)	17.406 (2.159)	10.204 (3.069)	6.574 (0.990)		
		4	31 (3.7)	20 (0.0)	7 (1.5)	7 (1.5)	6.354 (1.174)	40 (9.7)	20 (0.0)	4 (1.3)	3 (1.1)	20.252 (2.297)	12.564 (2.941)	7.744 (0.913)		

Table 2

Analysis of diabetes data.

Variable	Labels	M_1	M_2
age	Age of patient	0.0963	0.0117
height	Height of patient	0	0
weight	weight of patient	0	0
bp.1s	First Systolic Blood Pressure	0	0.0005
bp.1d	First Diastolic Blood Pressure	0	0
waist	waistline	0.0277	0.0063
hip	Hip measurement	0	0
time.ppn	Postprandial Time when Labs were Drawn	0.0192	0.0003
loca	Buckingham or Louisa	0	-0.0948
gend	Gender of patient	0	0
frame.m	Body frame size: medium	0.0003	0.0228
frame.l	Body frame size: large	0	0
stab.glu	stabilized glucose	–	0.0267
chol	total cholesterol	–	0.0037
hdl	high density lipoprotein	–	0
ratio	cholesterol/HDL ratio	–	0.0957