Electronic Journal of Statistics
Vol. 11 (2017) 3522–3548
ISSN: 1935-7524
DOI: 10.1214/17-EJS1329

BS-SIM: An effective variable selection method for high-dimensional single index model

Longjie Cheng

Department of Statistics, Purdue University
e-mail: longjiechengresearch@gmail.com

Peng Zeng

Department of Mathematics and Statistics, Auburn University
e-mail: zengpen@auburn.edu

and

Yu Zhu*

Department of Statistics, Purdue University
and
Center for Statistical Science, Department of Industrial Engineering, Tsinghua Unversity
e-mail: yuzhu@purdue.edu

Abstract: The single index model is an intuitive extension of the linear regression model. It has become increasingly popular due to its flexibility in modeling. Similar to the linear regression model, the set of predictors for the single index model can contain a large number of irrelevant variables. Therefore, it is important to select the relevant variables when fitting the single index model. However, the problem of variable selection for high-dimensional single index model is not well settled in the literature. In this work, we combine the idea of applying cubic B-splines for estimating the single index model with the idea of using the family of the smooth integration of counting and absolute deviation (SICA) penalty functions for variable selection. We propose a new method to simultaneously perform parameter estimation and model selection for the single index model. This method is referred to as the B-spline and SICA method for the single index model, or in short, BS-SIM. We develop a coordinate descent algorithm to efficiently implement BS-SIM. We also show that under certain conditions, the proposed method can consistently estimate the true index and select the true model. Simulations with various settings and a real data analysis are conducted to demonstrate the estimation accuracy, selection consistency and computational efficiency of BS-SIM.

MSC 2010 subject classifications: Primary 62H12; secondary 62G08.
Keywords and phrases: Single index model, variable selection, regression spline, LASSO, SICA.

Received June 2015.

*To whom correspondence should be addressed.

Contents

1. Introduction

Consider a univariate response Y and a p-dimensional predictor X. The single index model takes the following form

$$Y = f(X^T\theta_0) + \varepsilon, \tag{1}$$

where T indicates the transpose of a matrix, θ_0 is a vector of length p and referred to as the index, f is an unknown smooth function, and ε denotes the random error term. The single index model generalizes the linear model by incorporating a non-parametric link function f, and it has applications in a wide range of fields.

A number of methods have been proposed to estimate the true index θ_0 in the literature. Härdle and Stoker [5] introduced the Average Derivative Estimation (ADE) method. It relies on an intrinsic property of the single index model that θ_0 is proportional to the gradient $\partial f/\partial X$. Several modified ADE methods have been proposed later, including the density-weighted ADE method [16] and the out-product of gradients method [26]. Another category of estimation methods consist of methods that simultaneously estimate θ_0 and f. The Minimum Average Variance Estimation (MAVE) method proposed by Xia et al. [27] enjoys the most popularity among these methods. One major drawback of the ADE-based methods and the MAVE method is that they all use high dimensional kernels in estimation, and thus suffer from the curse of dimensionality. Consequently, they do not perform well in estimation even when the dimension p is moderate. To overcome this, Xia et al. [27] also proposed the refined MAVE (rMAVE) method by replacing the high dimensional kernel with a lower dimensional projection kernel. However, the computational complexity of MAVE and rMAVE still grows rapidly with the sample size n, and they can become unstable when

p increases. Recently, Wang and Yang [23] proposed the Single-Index Prediction (SIP) estimator by using the cubic B-splines to estimate θ_0 and f simultaneously. The application of the cubic B-splines circumvents the drawbacks suffered by high dimensional kernels, and as expected, simulation studies showed that SIP is considerably faster than MAVE, especially in the high dimensional case.

In practice, a large number of variables among the predictors may not be related to the response. Similar to the linear regression, it is important to select the relevant variables when fitting the single index model. Various traditional variable selection methods have been extended to the single index model; for example, AIC [11] and cross-validation [6]. However, these methods suffer from the same drawbacks as in the linear regression. They are intensive in terms of computation. Furthermore, it is infeasible to develop the large sample properties for the resulting estimates. Tibshirani [20] introduced the least absolute shrinkage and selection operator (LASSO) as a regularization method for simultaneous parameter estimation and variable selection in the linear models. LASSO has gained huge popularity since it was proposed, due to its succinctness and computational efficiency. Zhao and Yu [29] studied the sufficient and almost necessary condition, namely the Irrepresentable Condition, under which LASSO can consistently select the true model. Several attempts have been made to incorporate LASSO or its variants into the single index model; see [24, 14, 28, 25]. All of these methods combine some penalty function with MAVE, thus they inherit the drawbacks of MAVE. They are computationally inefficient for increasing sample size and become unstable when the dimensionality is high.

There are various extensions or variants of LASSO proposed in the literature; see [3, 31, 1] among others. Lv and Fan [9] considered a unified framework for regularized least squares methodology with a family of concave penalty functions. This family of penalty functions forms a smooth homotopy between the L_0 and L_1 penalities and thus is referred to as smooth integration of counting and absolute deviation (SICA) penalty functions. It includes LASSO as a limiting case. Lv and Fan [9] also developed the properties of the resulting estimator under the linear model and the SICA penalty function. More specifically, they obtained the conditions on the design matrix under which the estimator can recover the true model. These conditions on the design matrix are less restrictive than the Irrepresentable Condition, which may make the SICA penalty more appealing in cases where the Irrepresentable Condition does not hold and LASSO is not consistent in variable selection.

In this paper, we propose a new method to simultaneously perform parameter estimation and model selection for the single index model. This method combines the idea of using B-splines for estimating the single index model with the idea of using the SICA penalty for variable selection. We refer to it as the B-spline and SICA method for the single index model, or in short, BS-SIM. We develop a coordinate descent algorithm to efficiently implement our method for both low and high dimensionality. We prove that under mild regularity conditions, our method is consistent in estimation and can achieve the optimal estimation rate. We further show that with more conditions on the structure of f and the design matrix X, our method also has the ability to correctly

identify the true model. As mentioned in the previous paragraph, the LASSO penalty is a limiting case of the family of the SICA penalty functions. Therefore, the algorithm and the properties of BS-SIM are also applicable to LASSO. When LASSO is applied, the method is referred to as the B-spline and LASSO method for the single index model, or BL-SIM. The simulation studies and a real data example demonstrate that BS-SIM provides excellent computational efficiency, estimation accuracy and selection consistency for data of low to high dimensionality.

The rest of the paper is organized as follows. Section 2 describes BS-SIM and BL-SIM, and outlines an efficient algorithm to implement them. Section 3 presents the theoretical properties of the proposed methods. Section 4 and Section 5 illustrate the performance of the proposed methods in various simulation studies, as well as a real data example. The technical proofs are given in the Supplementary Material [32].

2. Methods

2.1. Spline estimation and regularization

Suppose a random sample of n observations is generated from the single index model

$$y_i = f(x_i^T\theta_0) + \varepsilon_i,$$

$i = 1, 2, \cdots, n$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \ldots, \theta_{0,p})^T$ is the true index, and ε_i's are i.i.d random variables with mean 0 and a common variance σ^2. Let $\mathbf{Y} = (y_1, \cdots, y_n)^T$ denote the $n \times 1$ reponse vector, and $\mathbf{X} = (x_1, x_2, \cdots, x_n)^T$ be the $n \times p$ matrix with x_i representing its i-th row. The true index θ_0 is only identifiable up to a scale constant without further constraint. In the literature, there are two popularly used identifiability constraints:

Identifiability Constraint 1: $\theta_{0,1} = 1$;
Identifiability Constraint 2: $\|\theta_0\|_2 = 1$ and $\theta_{0,1} > 0$.

In this work, we consider any general and feasible constraint on the scale of θ_0. We work with the nontrivial case that there is at least one non-zero component in θ_0. Thus, for any constraint, it is important to first identify one component $\theta_{0,k}$ that is non-zero. This component $\theta_{0,k}$ can be assumed as known or identified by methods such as marginal correlation. Without loss of generality, we assume $k = 1$. Although a large number of general identifiability constraints can be used, in Section 4, we show with simulation studies that different constraints can have different impacts on the performance of the used method in various aspects.

Suppose one specifies the following identifiability constraint: $\mathcal{C}(\theta) = 1$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$, and $\mathcal{C}$ is an explicit function on the scale of θ. Then θ_1 can be expressed as a function of the remaining components, that is, $\theta_1 = \mathcal{C}_1(\theta_2, \theta_3, \ldots, \theta_p)$. Let $\phi = (\theta_2, \theta_3, \ldots, \theta_p)^T$ be the $(p-1)$-dimensional sub-vector of θ by excluding the first component, and let $t_\theta = X^T\theta$. Let ϕ_0 denote the

last $(p-1)$ components of θ_0. Let Φ be the space for ϕ. With an appropriate identifiability constraint imposed, ϕ and θ have a one-to-one association. Then the goal of inference under the single index model is to estimate ϕ_0 (and thus θ_0) and the true link function f.

For a given θ, let $t_\theta^i = x_i^T\theta$ be the projected data onto the direction of θ, $i = 1, 2, \ldots, n$. Let $t_\theta(\min) = \min_i t_\theta^i$ and $t_\theta(\max) = \max_i t_\theta^i$. The interval $[t_\theta(\min), t_\theta(\max)]$ is partitioned into $(N+1)$ subintervals. Let T_N be the sequence of the N interior knots that separate the subintervals. Let $B_4 = (B_{4,1}, B_{4,2}, \ldots, B_{4,N+4})^T$ be the cubic B-spline basis functions on $[t_\theta(\min), t_\theta(\max)]$ with knots T_N. The explicit form of B_4 can be derived recursively [2]. Here we slightly abuse the notations in the sense that θ and T_N are omitted in the representation of the basis functions. The evaluations of the basis functions on the projected data points are denoted as $\mathbf{B}_\theta$. That is, $\mathbf{B}_\theta = (B_4(t_\theta^1), \ldots, B_4(t_\theta^n))^T$, where $B_4(t)$ denotes the evaluation of the cubic B-spline basis functions at t.

The cubic B-spline estimator of f is defined as $\hat{f}_\theta(\cdot) = \hat{\alpha}^T B_4(\cdot)$, where $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_{N+4})^T$, and can be obtained by solving the following least-squares problem

$$\min_{\alpha \in \mathbb{R}^{N+4}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \alpha^T B_4(t_\theta^i)\right)^2.$$

It immediately follows that $\hat{\alpha} = (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T \mathbf{Y}$. Note that $\hat{f}_\theta(\cdot)$ depends on θ. Wang and Yang [23] further proposed to use the following least-squares method to estimate θ_0

$$\hat{\theta}_{\text{un}} = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}_\theta(t_\theta^i))^2,$$

where $\hat{\theta}_{\text{un}}$ denotes the unpenalized estimator of θ_0, and $\Theta = \{\theta : \|\theta\|_2^2 = 1, \theta_1 > 0\}$. As discussed in Section 1, the dimension p can be high in practice, and the set of predictors can include a large number of irrelevant variables. Therefore, it is of interest to produce a sparse estimator of θ_0, and thus achieve automatic variable selection. This motivates us to utilize the spline estimator $\hat{f}_\theta(\cdot)$ for f described above, coupled with the regularized least squares method for estimating θ_0 to achieve efficient and simultaneous parameter estimation and variable selection.

Since $\theta_{0,1}$ is assumed to be non-zero, we penalize ϕ instead of θ. We further use the family of the SICA penalty functions. That leads us to the following objective function $R(\phi; \lambda)$.

$$R(\phi; \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}_\theta(t_\theta^i)\right)^2 + \lambda \sum_{j=1}^{p-1} \rho_a(|\phi_j|),$$

where $\hat{f}_\theta$ is the cubic B-spline estimator of f for a given θ, λ is a tuning parameter, and $\rho_a(u)$ denotes the SICA penalty function with the following form

$$\rho_a(u) = \left(\frac{u}{a+u}\right) I(u \neq 0) + \left(\frac{a}{a+u}\right) u, \qquad u \in [0, \infty),$$

where I is the indicator function. For simplicity, we do not include a in the notation of R, and write $R(\phi;\lambda)$ as $R(\phi)$ when there is no confusion. For a fixed λ, we define the following estimator of ϕ_0,

$$\hat{\phi} = \underset{\phi\in\Phi}{\operatorname{argmin}} R(\phi), \tag{2}$$

The corresponding estimator for θ_0 is denoted as $\hat{\theta}$, and is referred to as the BS-SIM estimator.

As discussed in [9], the SICA family of penalty functions provides a smooth homotopy between the L_0 and L_1 penalties, and we have

$$\rho_0(u) = \lim_{a\to 0+} \rho_a(u) = I(u \neq 0), \quad \text{and} \quad \rho_\infty(u) = \lim_{a\to\infty} \rho_a(u) = u.$$

That means, the LASSO penalty is the limiting case of the SICA penalty. In some applications, the LASSO penalty can also be of interest, and the estimator based on LASSO is defined separately below. We denote the objective function when $a = \infty$ as $R_L(\phi;\lambda)$. That is,

$$R_L(\phi;\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_\theta(t_\theta^i)\right)^2 + \lambda\|\phi\|_1,$$

where $\|\cdot\|_1$ denotes the L_1 norm. We write it as $R_L(\phi)$ when there is no confusion. For a fixed λ, we define the following estimator of ϕ_0,

$$\hat{\phi}^L = \underset{\phi\in\Phi}{\operatorname{argmin}} R_L(\phi), \tag{3}$$

and the corresponding estimator for θ_0 is denoted as $\hat{\theta}^L$. We refer to $\hat{\theta}^L$ as the BL-SIM estimator. It can be expected that the BS-SIM estimator can converge to the BL-SIM estimator as a approaches ∞.

2.2. Coordinate descent algorithm

For ease of representation, we define $H(\phi) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_\theta(t_\theta^i))^2$. Then the objective function $R(\phi)$ can be expressed as $R(\phi) = H(\phi) + \lambda\sum_{j=1}^{p-1}\rho_a(|\phi_j|)$. Next, we develop a coordinate descent algorithm to find $\hat{\phi}$ (or $\hat{\phi}^L$) for any given λ on a dense grid.

Since $H(\phi)$ is a complicated function of ϕ, we further use a local quadratic approximation strategy to iteratively solve Problem (2). Let $H^{(1)}(\cdot) = \partial H(\cdot)/\partial\phi$ and $H^{(2)}(\cdot) = \frac{\partial^2 H(\phi)}{\partial\phi\partial\phi^T}(\cdot)$, which are the gradient and Hessian matrix of H, respectively. Then, given a current estimate $\hat{\phi}^{(0)}$, the quadratic approximation to $H(\phi)$ at $\phi^{(0)}$ is given as follows.

$$\begin{aligned} H(\phi) &\approx H(\phi^{(0)}) + (\phi-\phi^{(0)})^T H^{(1)}(\phi^{(0)}) + \frac{1}{2}(\phi-\phi^{(0)})^T H^{(2)}(\phi^{(0)})(\phi-\phi^{(0)}) \\ &= \frac{1}{2}\phi^T H^{(2)}(\phi^{(0)})\phi - \phi^T\left(H^{(2)}(\phi^{(0)})\phi^{(0)} - H^{(1)}(\phi^{(0)})\right) + \text{constant}. \end{aligned} \tag{4}$$

In addition, we use a local approximation to the SICA penalty function suggested by [9] as follows.

$$\sum_{j=1}^{p-1} \rho_a(|\phi_j|) = \sum_{j=1}^{p-1} [\rho_a(|\phi_j^{(0)}|) + \rho_a'(|\phi_j^{(0)}|)(|\phi_j| - |\phi_j^{(0)}|)], \tag{5}$$

where $\phi^{(0)} = (\phi_1^{(0)}, \phi_2^{(0)}, \cdots, \phi_{p-1}^{(0)})^T$.

These two approximations entail that for a given $\phi^{(0)}$, Problem (2) can be approximated by

$$\min_{\phi \in \Phi} \frac{1}{2} \phi^T H^{(2)}(\phi^{(0)}) \phi - \phi^T \left(H^{(2)}(\phi^{(0)}) \phi^{(0)} - H^{(1)}(\phi^{(0)}) \right) + \lambda \sum_{j=1}^{p-1} w_j |\phi_j|, \tag{6}$$

where $w_j = \rho_a'(|\phi_j^{(0)}|)$ for $j = 1, 2, \cdots, p-1$. To solve Problem (6), we cyclically update each component of ϕ while holding the other components fixed. That means, for $j = 1, 2, \cdots, p-1$, we solve the following univariate problem

$$\min_{\phi_j} \frac{1}{2} h_{jj} \phi_j^2 + \left(\sum_{k=1, k \neq j}^{p-1} h_{jk} \phi_k - \beta_j \right) \phi_j + \lambda w_j |\phi_j| + \text{constant}, \tag{7}$$

where h_{kl} denotes the component in the kth row and the lth column of $H^{(2)}(\phi^{(0)})$, and β_j denotes the jth element of $H^{(2)}(\phi^{(0)})\phi^{(0)} - H^{(1)}(\phi^{(0)})$. Notice that Problem (7) is essentially a univariate LASSO problem, and the solution can be written down explicitly as

$$\phi_j = \text{sign}(a_j) \frac{(|a_j| - \lambda w_j)_+}{h_{jj}} = \begin{cases} (a_j - \lambda w_j)/h_{jj}, & \text{if } a_j > \lambda w_j; \\ (a_j + \lambda w_j)/h_{jj}, & \text{if } a_j < -\lambda w_j; \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

where $a_j = \beta_j - \sum_{k \neq j} h_{jk} \phi_k$. We repeatedly iterate through j and update the estimate of ϕ_0, until some convergence criterion is met.

When implementing Algorithm 1, there are two issues that require further attention. First, during the sth cycle of j, line search method is applied [12]. We start with $\hat{\phi}^{(s)}$, and obtain a tentative update $\hat{\phi}^{(s+1)}$. Before setting $\hat{\phi}^{(s+1)}$ as the most current estimate of ϕ_0, we need to check that the objective function R is indeed decreasing. If it is not, the step $\delta = \hat{\phi}^{(s+1)} - \hat{\phi}^{(s)}$ is repeatedly multiplied by 0.8, until the amount of movement along the direction δ that can result in a decrease in R is obtained. Here, 0.8 is chosen for the purpose of convenience, and may not be optimal. A more sophisticated choice can be further explored; see the previously mentioned reference on line search. The other issue faced during the implementation is that the optimization over ϕ should be carried out in the space Φ. However, the algorithm described above does not consider any constraint on the space over which the optimization is executed. For some identifiability constraints, such as the *Identifiability Constraint 1* mentioned earlier,

Φ is actually $\mathbb{R}^{p-1}$; for other identifiability constraints, such as the *Identifiability Constraint 2* in the previous section, Φ is a constrained subspace of $\mathbb{R}^{p-1}$. In the former case, no adjustment is needed; in the latter case, there requires an additional step that ensures that the updated $\hat{\phi}$ is in the constrained space Φ. For instance, it needs to be checked that the updated $\hat{\phi}$ satisfies $\|\hat{\phi}\|_2 < 1$, for *Identifiability Constraint 2*. If it does not, the step δ needs to be shortened such that $\hat{\phi}$ falls within Φ. Algorithm 1 outlines the search for $\hat{\phi}$ at a given λ in more detail. Problem (3) can be solved in a similar fashion. The only difference is that for Problem (3), there is no need to use the local linear approximation to the penalty function. Therefore, the algorithm of searching for $\hat{\phi}^L$ is not separately displayed.

Algorithm 1 *Coordinate Descent Algorithm*

For any λ,

1. Initialize ϕ to be $\hat{\phi}^{(0)}$ and let $s = 0$.
2. Given $\hat{\phi}^{(s)} = (\hat{\phi}_1^{(s)}, \hat{\phi}_2^{(s)}, \ldots, \hat{\phi}_{p-1}^{(s)})^T$, calculate the quadratic approximation (4) to $H(\phi)$ and the linear approximation (5) to $p_\lambda(\phi)$.
3. For $j = 1, 2, \ldots, p-1$, update $\hat{\phi}_j$ by the following formulars:
$$\phi_j = \text{sign}(a_j)\frac{(|a_j| - \lambda w_j)_+}{h_{jj}} = \begin{cases} (a_j - \lambda w_j)/h_{jj}, & \text{if } a_j > \lambda w_j; \\ (a_j + \lambda w_j)/h_{jj}, & \text{if } a_j < -\lambda w_j; \\ 0, & \text{otherwise.} \end{cases}$$
If needed, check whether ϕ is within Φ. If it is not, adjust it to fall within Φ.
4. After one cycle of j, a tentative update $\hat{\phi}^{(s+1)}$ and the corresponding $R(\hat{\phi}^{(s+1)})$ are obtained. If $R(\hat{\phi}^{(s+1)}) > R(\hat{\phi}^{(s)})$, calculate $\delta = \hat{\phi}^{(s+1)} - \hat{\phi}^{(s)}$, and check the objective function for
$$\hat{\phi}^{(s+1)} = \hat{\phi}^{(s)} + (0.8)^k \delta,$$
for $k = 1, 2, \ldots$ until $R(\hat{\phi}^{(s+1)})$ is smaller than $R(\hat{\phi}^{(s)})$.
5. Calculate $\Delta = R(\hat{\phi}^{(s)}) - R(\hat{\phi}^{(s+1)})$. If Δ is below a prespecified threshold, then stop and set $\hat{\phi} = \hat{\phi}^{(s+1)}$ and calculate the corresponding $\hat{\theta}$; otherwise, set $s = s + 1$ and go back to Step 2.

2.3. Tuning parameter selection

For regularization-based approaches, it is crucial to choose the tuning parameters, namely λ and a in our case. We start with the discussion of the selection of λ. We consider two types of methods for determining λ. The first one is m-fold cross-validation, denoted as CV hereafter. In CV, the sample is randomly partitioned into m subsamples of equal size. Among these m folds, $m-1$ of them are treated as the training set, and the remaining one is treated as the validation set. At each given candidate value for λ, the proposed approach is applied to the training set, and a fitting is obtained. Subsequently, the test set is used to assess the predictive accuracy of the obtained model. The residual sum of squares can be used as the assessment. This process is repeated m times until each fold of the sample is used as the test set exactly once. For a given λ,

the m results on the assessment are then averaged. The value of λ that yields the smallest average is regarded as optimal.

The second type is the Bayesian Information Criterion (BIC) and its variants [19]. For variable selection under the linear model $Y = X^T\beta + \epsilon$, we examine the following four BIC-based criteria (9)-(12).

$$\text{BIC} = \text{RSS}_\lambda/n + d\sigma^2\log(n)/n, \tag{9}$$

$$\text{logBIC} = \log(\text{RSS}_\lambda/n) + d\log(n)/n, \tag{10}$$

$$\text{GIC} = \text{RSS}_\lambda/n + d\sigma^2 k_n/n, \tag{11}$$

$$\text{logGIC} = \log(\text{RSS}_\lambda/n) + dk_n/n, \tag{12}$$

where RSS_λ denotes the residual sum of squares at a given λ, σ^2 denotes the error variance, and d is the size of the identified model at a given λ. Furthermore, for criteria (11) and (12), k_n represents the additional penalty imposed on the size of the model. In practice, σ^2 is rarely known. On the other hand, according to [18], under certain conditions, the BIC defined in (9) has the same asymptotic behavior as the one defined in (10). Thus, it is more convenient to rely on logBIC in (10) to select the tuning parameter λ. It has been previously proved that, when the number of predictors p is fixed as the number of observations n grows, one can identify the true model with probability tending to 1 in the linear models by using the logBIC criteria [22]. Nevertheless, when p diverges, the logBIC criterion (10) tends to yield a model that contains many irrelevant predictors. Several adjustments have been proposed in the literature to circumvent this issue [22, 4]. The common approach these adjustments take is to place more penalty on the model complexity d. This idea naturally leads us to consider the GIC criterion in (11) and the logGIC criterion in (12). It is clear that GIC and logGIC include BIC and logBIC as a special case, respectively. Thus, GIC and logGIC can be regarded as the unified criteria to achieve the selection of λ for any p, and they can be extended to models other than the linear regression models. It is also worth noting that GIC involves σ^2. When σ^2 is unknown, there are various ways to obtain an estimate $\hat{\sigma}^2$ and replace σ^2 with $\hat{\sigma}^2$ in GIC. We will elaborate on it in the next paragraph.

In order to choose a proper type of method for determining λ under our framework, we carry out extensive simulation studies under both the linear model and the single index model. We try different settings of p and the size of the true model. In the simulation studies, we use $\hat{\sigma}^2 = \text{RSS}_0/(n-p)$ when $n > p$, and $\hat{\sigma}^2 = \text{RSS}_{\lambda_{cv}}/(n - d_{cv})$ otherwise, where λ_{cv} denotes the value of λ selected by CV, and d_{cv} denotes the size of the model selected by CV. For all settings, CV generally leads to an overfitted model. When the true model is sparse, logGIC with an appropriate k_n performs the best in terms of identifying the true model for any p. GIC is a close second. As the number of relevant variables grows, the performance of GIC surpasses that of logGIC, and GIC becomes the most preferable. For the moderately sparse scenario, logGIC starts to break down as p increases. When the size of the true model is large, logGIC fails to work in the sense that it leads to either a very large model, or a very small model. Meanwhile, GIC can still produce significant improvement over

CV when p is not large. When p also becomes large, the problem itself becomes too difficult that all of the methods rarely perform satisfactorily.

Based upon these observations, we propose the following rule of thumb principle for the selection of λ under our framework. When sparsity of the true model is assumed, we use logGIC; when the size of the true model is relatively large, we use GIC. An example illustrating the breakdown of logGIC and the advantage of using GIC under the violation of the sparsity assumption is given in Section 4 of the Supplementary Material [32].

As for the selection of a, it can generally be accomplished by m-fold cross-validation. Since the focus of this work is to study the properties of $\hat{\theta}$ and $\hat{\theta}^L$, we do not intensively examine the selection of a.

3. Theoretical properties

Before stating the theoretical properties of BS-SIM and BL-SIM, we first need to impose the following regularity conditions (A1)-(A3).

(A1) The link function f has continuous and bounded second order derivative.

(A2) Let $R^*(\phi) = E[Y - f(X^T\theta)]^2$ be the population risk function. Define $H^{*(2)}(\phi) = \frac{\partial^2 R^*(\phi)}{\partial\phi\partial\phi^T}$ as the Hessian matrix of $R^*(\phi)$. $H^{*(2)}(\phi_0)$ is positive definite, and its smallest eigenvalue is $\rho(\min)$, for some $\rho(\min) > 0$.

(A3) The number of interior knots N satisfies $N \sim n^{1/5}$.

3.1. Estimation consistency

To begin with, we show that, under mild conditions, $\hat{\theta}$ is consistent in terms of estimation, and can achieve the optimal $\sqrt{n}$ rate for a well-selected λ. Moreover, as a special case, $\hat{\theta}^L$ share the same property on parameter estimation. We will also show the theoretical property of $\hat{f}$ in terms of estimating f.

Theorem 1. *Suppose the regularity conditions (A1)–(A3) hold.*

(a) If $\lambda = O(n^{-1/2})$, there exists a local minimum $\hat{\phi}$ of $R(\phi)$, such that $\hat{\phi}$ is $\sqrt{n}$-consistent. Consequently, the BS-SIM estimator $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ_0;

(b) If $\lambda = O(n^{-1/2+\delta})$ for some $\delta \in (0, 1/2)$, there exists a local minimum $\hat{\phi}$ of $R(\phi)$, such that $\|\hat{\phi} - \phi_0\|_2 = O_p(n^{-1/2+\delta})$. As a result, $\|\hat{\theta} - \theta_0\|_2 = O_p(n^{-1/2+\delta})$;

(c) As a special case, the BL-SIM estimator $\hat{\theta}^L$ possesses the above properties.

Theorem 1 is expected and standard. Part (b) of Theorem 1 also facilitates the derivations on the selection consistency given in the next subsection. The next theorem characterizes the convergence rate of $\hat{f}$ as an estimator of the regression function f.

Theorem 2. *Suppose Conditions (A1)–(A3) hold. If* $\lambda = O(n^{-1/2+\delta})$ *for some* $\delta \in [0, 1/2)$*, there exists a local minimum* $\hat{\phi}$ *of* $R(\phi)$ *such that*

$$\left\| f(t_{\theta_0}) - \hat{f}_{\hat{\theta}}(t_{\hat{\theta}}) \right\|_{l_2} = O_p\left((nh)^{-1/2} logn + h^4 + n^{-1/2+\delta/2} (logn)^{1/2} \right),$$

where $\left\| f(t_{\theta_0}) - \hat{f}_{\hat{\theta}}(t_{\hat{\theta}}) \right\|_{l_2}^2 = \frac{1}{n} \sum_{i=1}^n \left(f(t^i_{\theta_0}) - \hat{f}_{\hat{\theta}}(t^i_{\hat{\theta}}) \right)^2$.

3.2. Intuition and notations for selection consistency

Observe that if no identifiability constraint is imposed, we have $f(t_\theta) - f(t_{\theta_0}) \approx D'_\theta(t_{\theta_0})(\theta - \theta_0)$, where $D'_\theta(t_{\theta_0}) = \left(\frac{\partial f(t_{\theta_0})}{\partial \theta_1}, \frac{\partial f(t_{\theta_0})}{\partial \theta_2}, \ldots, \frac{\partial f(t_{\theta_0})}{\partial \theta_p} \right)$. By simple calculations, we obtain

$$\frac{\partial f(t^i_{\theta_0})}{\partial \theta_j} = h(t^i_{\theta_0}) X_{ij} \triangleq g_{ij},$$

where $h(t^i_{\theta_0}) = f'|_{t=t^i_{\theta_0}}$ for $j = 1, 2, \ldots, p$, and $i = 1, 2, \ldots, n$. Let

$$F = \begin{pmatrix} \frac{\partial f(t^1_{\theta_0})}{\partial \theta_1}, & \frac{\partial f(t^1_{\theta_0})}{\partial \theta_2}, & \cdots, & \frac{\partial f(t^1_{\theta_0})}{\partial \theta_p} \\ \frac{\partial f(t^2_{\theta_0})}{\partial \theta_1}, & \frac{\partial f(t^2_{\theta_0})}{\partial \theta_2}, & \cdots, & \frac{\partial f(t^2_{\theta_0})}{\partial \theta_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(t^n_{\theta_0})}{\partial \theta_1}, & \frac{\partial f(t^n_{\theta_0})}{\partial \theta_2}, & \cdots, & \frac{\partial f(t^n_{\theta_0})}{\partial \theta_p} \end{pmatrix}_{n \times p} = (g_{ij})_{i=1,2,\ldots,n; j=1,2,\ldots,p}.$$

By the definition of g_{ij}, it is apparent that F is a weighted design matrix. That is, F is computed by multiplying row i of X with the corresponding derivative of f at $t^i_{\theta_0}$, $h(t^i_{\theta_0})$, for $i = 1, 2, \cdots, n$. When f is flat at $t^i_{\theta_0}$, this data point does not contain much information on θ_0, and the weight placed on row i is small; on the other hand, when f is steep at $t^i_{\theta_0}$, this data point is informative, and the corresponding row is scaled with a larger weight. In the special case of the linear models, F reduces to X.

However, θ_0 is not free of identifiability constraint, and only the last $p-1$ elements of θ_0 are of interest. Consequently, we consider

$$F_0 = \begin{pmatrix} \frac{\partial f(t^1_{\theta_0})}{\partial \theta_2}, & \frac{\partial f(t^1_{\theta_0})}{\partial \theta_3}, & \cdots, & \frac{\partial f(t^1_{\theta_0})}{\partial \theta_p} \\ \frac{\partial f(t^2_{\theta_0})}{\partial \theta_2}, & \frac{\partial f(t^2_{\theta_0})}{\partial \theta_3}, & \cdots, & \frac{\partial f(t^2_{\theta_0})}{\partial \theta_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(t^n_{\theta_0})}{\partial \theta_2}, & \frac{\partial f(t^n_{\theta_0})}{\partial \theta_3}, & \cdots, & \frac{\partial f(t^n_{\theta_0})}{\partial \theta_p} \end{pmatrix}_{n \times (p-1)}.$$

Here, F_0 depends on the design X, the true link function f, and the true index θ_0. To some extent, F_0 can be treated as the design matrix in the single index models, and it can play a crucial role in the subsequent analysis. For a given

identifiability constraint, we can express θ_1 as a function of the rest $(p-1)$ components of θ, that is $\theta_1 = \mathcal{C}_1(\theta_2, \ldots, \theta_p)$. Let J be the corresponding Jacobian matrix for θ_0, that is,

$$J = \begin{pmatrix} \frac{\partial \mathcal{C}_1(\phi_0)}{\partial \theta_2}, & \frac{\partial \mathcal{C}_1(\phi_0)}{\partial \theta_3}, & \cdots & \frac{\partial \mathcal{C}_1(\phi_0)}{\partial \theta_p} \\ 1, & 0, & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0, & 0, & \cdots & 1 \end{pmatrix}_{p\times(p-1)}.$$

And it follows that $F_0 = FJ$. For simplicity, here we omit the dependence of J on the identifiability constraint in the notation. The forms of F_0 for the two popular identifiability constraints are illustrated in Section 1 of the Supplementary Material [32]. Notice that F_0 is essentially a scaled and adjusted version of the design matrix X.

Without the loss of generality, let $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \ldots, \theta_{0,q}, \theta_{0,q+1}, \ldots, \theta_{0,p})^T$ where $\theta_{0,j} \neq 0$ for $j = 1, 2, \ldots, q$ and $\theta_{0,j} = 0$ for $j = q+1, q+2, \ldots, p$. Let $\mathcal{A}_1 = \{2, 3, \ldots, q\}$ and $\mathcal{A}_2 = \{q+1, q+2, \ldots, p\}$. For any ϕ, we also decompose it into two sub-vectors as follows $\phi(1) = (\theta_2, \theta_3, \ldots, \theta_q)^T$, and $\phi(2) = (\theta_{q+1}, \ldots, \theta_p)^T$. Let $C_0 = \frac{1}{n} F_0^T F_0$. Let $F_0(1)$ and $F_0(2)$ be the first $q-1$ and the last $p-q$ columns of F_0. Let $C_0(11) = \frac{1}{n} F_0^T(1) F_0(1)$, $C_0(21) = \frac{1}{n} F_0^T(2) F_0(1)$, $C_0(12) = \frac{1}{n} F_0^T(1) F_0(2)$ and $C_0(22) = \frac{1}{n} F_0^T(2) F_0(2)$. Then we can decompose C_0 into the following four blocks

$$C_0 = \begin{pmatrix} C_0(11) & C_0(12) \\ C_0(21) & C_0(22) \end{pmatrix}.$$

In the following subsections, we also rely on this decomposition to formulate the results on the selection consistency of the proposed estimators.

3.3. Selection consistency

As detailed earlier, we use the cubic spline function to estimate the true link function f. For any θ, let $\Gamma(\theta)$ be the cubic spline space defined according to Section 2.1. We denote the projection matrix onto $\Gamma(\theta)$ as $\mathbf{P}_\theta = \mathbf{B}_\theta (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T$. Thus,

$$\hat{f}_\theta = \left(\hat{f}_\theta(t_{\theta,1}), \ldots, \hat{f}_\theta(t_{\theta,n}) \right)^T = \mathbf{P}_\theta \mathbf{Y}.$$

Consequently, we have

$$\mathbb{E}\left(\hat{f}_\theta(t_\theta^i) \right) = \mathbf{P}_\theta f(t_{\theta_0}^i) \triangleq \bar{f}_\theta(t_\theta^i),$$

for $i = 1, 2, \cdots, n$. Then, for any given θ, we can similarly define $\bar{F}_\theta$ and $\bar{C}_\theta$ as

$$\bar{F}_\theta = \left(\frac{\partial \bar{f}_\theta(t_\theta^i)}{\partial \theta_j} \right)_{i=1,2,\ldots,n; j=2,3,\ldots,p},$$

and $\bar{C}_\theta = \frac{1}{n}\bar{F}_\theta^T \bar{F}_\theta$. For succinctness, we write $\bar{F}_{\theta_0}$ and $\bar{C}_{\theta_0}$ as $\bar{F}_0$ and $\bar{C}_0$. Different from F_0, $\bar{F}_0$ not only depends on X, f and θ_0, it also relies on the spline approximation of the link function. We decompose $\bar{C}_0$ into four blocks in the same way we decompose C_0. With the notations introduced above, we can impose the following crucial conditions on $\bar{C}_0$ to establish the selection consistency of BS-SIM.

Condition 1 (Irrepresentable Conditions for BS-SIM)**.** *$\bar{C}_0$ satisfies that*

$$\begin{aligned} \|\bar{C}_0^{-1}(11)\|_\infty &\leqslant \bar{L}_1, \\ \|\bar{C}_0(21)\bar{C}_0^{-1}(11)\|_\infty &\leqslant \bar{L}_2, \end{aligned}$$

where $\bar{L}_1 \in (0,\infty)$, $\bar{L}_2 \in \left(0, \bar{L}\frac{\rho'(0+)}{\rho'(b_0-\lambda\bar{L}_3)}\right)$ for some $\bar{L}$ and $\bar{L}_3 \in (0,\infty)$, and $b_0 = \min_{j\in\mathcal{A}_1} |\theta_{0,j}|$.

Note that $\bar{C}_0$ is related to the spline estimator of f, and thus it depends on the number and the location of the knots. That means the conditions given above are not free of the sample size n. On the other hand, $\bar{F}_0$ is a scaled and adjusted version of the design matrix X. Hence, the Irrepresentable Conditions for BS-SIM are similar to the conditions by [9] in the sense that the above conditions replace the design matrix X in [9] with $\bar{F}_0$. With the Irrepresentable Conditions for BS-SIM, we are ready to state our theorem next.

Theorem 3. *Assume the Irrepresentable Conditions for BS-SIM hold, and the regularity conditions (A1)-(A3) are satisfied. Then for $\lambda = O(n^{c-2/5})$, with some $c \in (0, 2/5)$, there exists a local minimum $\hat{\phi}$ of $R(\phi)$ such that*

$$P\left(sign(\hat{\phi}) = sign(\phi_0)\right) = 1 - o(e^{-n^c}), \text{ as } n \to \infty,$$

where $sign(s)$ is the sign function that equals 1 when s is positive, equals -1 when s is negative, and equals 0 when $s = 0$.

Theorem 3 characterizes the behaviour of BS-SIM in recovering the true model. It suggests that, if the Irrepresentable Conditions for BS-SIM hold, then the probability that BS-SIM is able to identify the true model converges to 1 exponentially. It can be easily shown that $\rho'(0+) = 1 + a^{-1}$. As noted by [9], the conditions for SICA to identify the true model in the linear regression becomes less restrictive as a decreases, at the sacrifice of computational convenience. This statement also holds in the context of the single index model. That means, with smaller a, the Irrepresentable Conditions for BS-SIM are less restrictive, but it is harder to find $\hat{\phi}$. As pointed out earlier, LASSO is a limiting case of the SICA penalty. Therefore, it is expected that the BL-SIM estimator $\hat{\phi}^L$ would possess the similar properties as given in Theorem 3. To present the properties for $\hat{\phi}^L$, we start with the following assumption on $\bar{C}_0$.

Condition 2 (Irrepresentable Condition for BL-SIM)**.** *There exists a positive constant vector $\bar{\eta}$, such that the following inequality holds component-wise*

$$\left|\bar{C}_0(21)\bar{C}_0^{-1}(11)sign(\phi_0(1))\right| \leqslant \mathbf{1}_{p-q} - \bar{\eta},$$

where $\mathbf{1}_{p-q}$ *denotes a vector of* 1*'s of length* $p-q$.

Again, the Irrepresentable Condition for BL-SIM resembles the Irrepresentable Condition in [29], and the major difference is that the Irrepresentable Condition for BL-SIM replaces X with $\bar{F}_0$.

Theorem 4. *Assume the Irrepresentable Condition for BL-SIM holds, and the regularity conditions (A1)-(A3) are satisfied. Then for* $\lambda = O(n^{c-2/5})$*, with some* $c \in (0, 2/5)$*, there exists a local minimum* $\hat{\phi}^L$ *of* $R_L(\phi)$ *such that*

$$P\left(sign(\hat{\phi}^L) = sign(\phi_0)\right) = 1 - o(e^{-n^c}).$$

Theorem 4 demonstrates that with the Irrepresentable Condition for BL-SIM imposed, the probability that BL-SIM selects the true model approaches 1 exponentially. Consistent with the monotonicity of the restrictiveness of the conditions, the Irrepresentable Condition for BL-SIM is more restrictive than the Irrepresentable Conditions for BS-SIM with finite a. This observation is also in line with that in the linear regression scenario, and it implies that BS-SIM may be able to recover the true model when BL-SIM fails.

Recall that the conditions presented previously rely on the sample size n. In what follows, we show that if $\bar{C}_0$ satisfies certain regularity condition, the selection consistency of the proposed methods can be achieved under conditions that are independent of n. From [23], we have

$$\sup_{j=2,3,\ldots,p} \sup_{\theta:\|\theta\|_2=1} \max_i \left|\frac{\partial}{\partial\theta_j}(\bar{f}_\theta - f)(t_\theta^i)\right| = O\left(h^3\right),$$

where $h = 1/(N+1)$ is the bandwith for the cubic B-spline functions. This means that $(\bar{F}_0)_i \to (F_0)_i$, as $n \to \infty$, for any i, and $(\cdot)_i$ denotes the ith row of a matrix. Based on this result, the following regularity condition can be imposed,

$$\bar{C}_0 \to C, \text{ as } n \to \infty,$$

for some matrix C free of n. We decompose C into four blocks in the same way we decompose C_0. Next, we show that if the Irrepresentable Conditions on C are imposed, the proposed methods can consistently select the true variables.

Condition 3 (Limiting Irrepresentable Conditions for BS-SIM)**.** *C satisfies that*

$$\begin{aligned} \|C^{-1}(11)\|_\infty &\leqslant L_1, \\ \|C(21)C^{-1}(11)\|_\infty &\leqslant L_2, \end{aligned}$$

where $L_1 \in (0, \infty)$*, and* $L_2 \in \left(0, L\frac{\rho'(0+)}{\rho'(b_0 - \lambda L_3)}\right)$ *for some* L *and* $L_3 \in (0, \infty)$.

Condition 4 (Limiting Irrepresentable Condition for BL-SIM)**.** *There exists a positive constant vector η, such that the following inequality holds componentwise*

$$\left|C(21)C^{-1}(11)sign(\phi_0(1))\right| \leqslant \mathbf{1}_{p-q} - \eta,$$

where $\mathbf{1}_{p-q}$ denotes a vector of 1's of length $p-q$.

Corollary 5. *(a) Assume that λ satisfies that $\lambda \sim n^{c-2/5}$, for some $c \in (0, 2/5)$, and the Limiting Irrepresentable Conditions for BS-SIM hold. Under regularity conditions (A1)-(A3), there exists a local minimum $\hat{\phi}$ of $R(\phi)$ such that*

$$P\left(sign(\hat{\phi}) = sign(\phi_0)\right) = 1 - o(e^{-n^c}).$$

(b) Assume that λ satisfies that $\lambda \sim n^{c-2/5}$, for some $c \in (0, 2/5)$, and the Limiting Irrepresentable Condition for BL-SIM holds. Under regularity conditions (A1)-(A3), there exists a local minimum $\hat{\phi}^L$ of $R_L(\phi)$ such that

$$P\left(sign(\hat{\phi}^L) = sign(\phi_0)\right) = 1 - o(e^{-n^c}).$$

Corollary 5 suggests that under the corresponding Limiting Irrepresentable Conditions, BS-SIM and BL-SIM can consistently recover the true model. On the other hand, same as the statements given in the last subsection, the Limiting Irrepresentable Conditions for BS-SIM become less restrictive as a decreases. As a result, the Limiting Irrepresentable Condition for BL-SIM is more restrictive than those for BS-SIM with finite a. The proofs of the theorems and the corollaries can be found in Section 2 of the Supplementary Material [32].

Remark 1. In the technical proofs provided in Section 2 of the Supplementary Material [32], we use the identifiability constraint that $\|\theta_0\|_2 = 1$ and $\theta_{0,1} > 0$. Nevertheless, the properties presented above should hold for any reasonable constraint, and one should be able to derive the proof for other conditions without much difficulty.

4. Simulation results

In this section, we present the results from five simulation studies. We demonstrate that the proposed regularization approach used is indeed beneficial in several aspects. We also look at the impact of the tuning parameter a on the performance of the resulting estimator, and point out a reasonable choice of a in practice. Subsequently, we compare the performance of the proposed methods to other existing methods for moderate to large p. For the comparison when p is small, we refer to Section 3 of the Supplementary Material [32]. The last simulation example is concerned about the impact that the Irrepresentable Condition has on our proposed method's ability of recovering the true model. For the purpose of succinctness, we use V1 and V2 to denote the *Identifiability Constraint 1* and *Identifiability Constraint 2* in this section, respectively. For the link function, we consider the following three models:

1. $Y = X^T\theta_0 + 4\sqrt{|X^T\theta_0 + 1|} + \varepsilon$;
2. $Y = 1 + 2(X^T\theta_0 + 3)\log(3|X^T\theta_0| + 1) + \varepsilon$;
3. $Y = (X^T\theta_0)^2 + \varepsilon$.

The models above are refered to as Model 1, Model 2, and Model 3, respectively. Furthermore, let Σ be a p-by-p matrix with the diagonal elements equal 1 and the off-diagonal element in kth row and lth column equal ρ_{kl}. Each x_i is sampled from $N(\mathbf{0}, \Sigma)$. The errors ε_i's are independently sampled from $N(0, 1)$. We examine the following three forms of Σ:

1. (No correlation) $\rho_{kl} = 0$, for $k \neq l$;
2. (Constant correlation) $\rho_{kl} = 0.3$, for $k \neq l$;
3. (Decaying correlation) $\rho_{kl} = 0.5^{|k-l|}$, for $k \neq l$.

We denote these three types of correlation structure as COR1, COR2, COR3, respectively.

For the first four examples, four metrics are used to assess the performance of an estimator, which are Angle, False Positive Rate (FPR), Ture Positive Rate (TPR) and Computing Time (Time), respectively. Angle is defined as $\text{Angle} = \arccos(\theta_0^T\hat{\theta})$, where θ_0 is the true index and $\hat{\theta}$ is an estimate, and they are standardized to have unit norm. FPR is defined as the ratio of the number of falsely identified predictors to the total number of identified predictor. TPR is the ratio of the number of correctly identified predictors to the total number of true relevant predictors. Finally, Time is the average time (in seconds) needed to obtain the estimate for one data set. In Examples 2–4, we search the best estimate on a dense grid of λ, and thus, Time represents the total amount of time consumed to find the estimate on the whole grid and yield the final estimate. On the other hand, in Example 1, Time refers to the amount of time used to find the estimate for a particular λ. In the tables presented in this section, the best performance on each metric is highlighted.

Example 1. This example compares the performance of the proposed estimator to that of the unpenalized estimator. We consider a moderate dimension $p = 70$ with $q = 8$ and $\theta_0 = (2.0, -1.0, 0.5, 1.0, -1.5, 1.0, -0.3, 1.2, 0, \cdots, 0)^T$. 100 samples of size $n = 100$ are generated from Model 1 with COR1. The coordinate descent algorithm described in Section 2.2 is used to implement BS-SIM with $a = 0.1$. The tuning parameter λ is chosen by three criteria, denoted as logBIC, logGIC1, and logGIC2, respectively. They correspond to three choices of k_n for logGIC defined in Section 2.3, which are $k_n^0 = \log(n)$, $k_n^1 = \text{loglog}n\log p$, and $k_n^2 = \log p\sqrt{\log n}$, respectively. Our method with $\lambda = 0$ is also applied to obtain the unpenalized estimate for θ_0. In this example, only V2 is used.

Table 1 shows the comparison results on the four aforementioned assessments. In terms of estimation accuracy and computing efficiency, both the BL-SIM estimators and the BS-SIM estimators are considerably better than the unpenalized estimator. It is a strong sign that the proposed regularization approach substantially helps with efficiently providing a more accurate estimator. Comparing the two proposed estimators, the BS-SIM estimators slightly outperform the BL-

Model 1, COR1, $p = 70$					
Method	Selection of λ	Angle	FPR	TPR	Time
BS-SIM-V2	logBIC	4.836 (2.309)	0.124	0.984	0.781
	logGIC1	4.529 (1.868)	0.050	0.976	0.701
	logGIC2	**4.526** (1.976)	**0.015**	0.968	0.610
BL-SIM-V2	logBIC	7.010 (5.421)	0.466	**0.995**	0.796
	logGIC1	6.828 (4.178)	0.457	**0.995**	0.577
	logGIC2	6.228 (3.114)	0.428	0.975	**0.453**
Unpenalized	$\lambda = 0$	50.350 (7.587)	NA	NA	12.749

TABLE 1

Comparison between the penalized estimator and the unpenalized estimator.

SIM estimators in estimation. In terms of the performance on variable selection consistency, the BS-SIM estimators are dramatically better. More specifically, the BL-SIM estimators have a more than 3-fold higher average FPR, indicating applying LASSO is more likely to lead to an overfitted model. In the computational efficiency aspect, BS-SIM is slightly faster than BL-SIM. As for the comparison among the three BS-SIM estimators, the estimator using logBIC has a noticeably higher average FPR than the estimators with λ chosen by logGIC1 and logGIC2. Since the number of predictors is not that small ($p = 70$) in this example, this observation on FPR is consistent with the fact that logBIC yields a overfitted model when the dimension p increases. The performance of the two penalized estimators with λ chosen by logGIC1 and logGIC2 are similar in terms of the four metrics.

Example 2. This example examines the performance of the proposed estimator for several choices of a. 100 samples of size 100 are simulated from Model 2 with COR1. The other settings are $\theta_0 = (2.0, -1.0, 0.5, 1.0, -1.5, 1.0, -0.3, 1.2, 0, \cdots, 0)^T$ and $p = 50$, $q = 8$. BL-SIM and BS-SIM with several choices of a are applied, and their performance on the four assessments introduced previously is compared. We rely on both logBIC and logGIC2 defined in Example 1 to choose the tuning parameter λ, and only use V2 in this example.

The comparison results are shown in Table 2. It can be observed that as a increases, both Angle and FPR decrease first, then increase. Furthermore, when a continues to increase, the performance of the BS-SIM estimator approaches that of the BL-SIM estimator. In theory, the performance of the BS-SIM estimator in terms of variable selection should improve when a decreases. Nevertheless, the pattern shown in Table 2 implies that there exists certain computational difficulty in finding a consistent estimate when a is extremely small. On the other hand, the BS-SIM estimator with $a = 0.1$ outforms the rest in terms of selection consistency. When it comes to estimation accuracy, the performance of the BS-SIM estimator with $a = 0.1$ is also satisfactory. Therefore, we recommend to use $a = 0.1$ in practice. For the remaining examples, we fix a at 0.1, unless otherwise specified.

Example 3. This example illustrates the performance of the proposed estimator for moderate p. We focus on the comparison between our method and other existing methods. In this example, we implement the proposed BS-SIM method

Model 2, COR1, $p = 50$					
Method	Selection of λ	Angle	FPR	TPR	Time
BS-SIM-V2	logBIC	1.392 (0.578)	0.075	1	**26.00**
($a = 0.01$)	logGIC2	1.160 (0.440)	0.013	1	
BS-SIM-V2	logBIC	1.178 (0.396)	0.029	1	32.78
($a = 0.05$)	logGIC2	**1.122** (0.381)	0.005	1	
BS-SIM-V2	logBIC	1.197 (0.399)	0.029	1	38.65
($a = 0.10$)	logGIC2	1.164 (0.397)	**0.004**	1	
BS-SIM-V2	logBIC	1.503 (0.468)	0.140	1	77.70
($a = 0.50$)	logGIC2	1.504 (0.474)	0.132	1	
BS-SIM-V2	logBIC	1.639 (0.472)	0.384	1	103.97
($a = 1.00$)	logGIC2	1.630 (0.470)	0.383	1	
BL-SIM-V2	logBIC	1.938 (0.557)	0.417	1	103.63
($a = \infty$)	logGIC2	1.925 (0.541)	0.413	1	

TABLE 2

Comparison between the LASSO and the SICA penalties with various choices of a for moderate p.

Model 1, COR2, $p = 50$				
Method	Angle	FPR	TPR	Time
BS-SIM-V1	4.866 (2.850)	0.019	0.963	34.463
BS-SIM-V2	**4.819** (2.749)	**0.017**	0.963	**25.262**
BL-SIM-V1	13.269 (3.956)	0.347	0.963	64.532
BL-SIM-V2	8.626 (3.121)	0.160	0.968	52.522
SIM-LASSO-V2	7.476 (2.085)	0.552	**0.990**	56.845
SMAVE-V2	12.493 (9.445)	0.316	0.898	39.747
SIM-Bridge-V2	7.686 (4.434)	0.058	0.901	102.349

TABLE 3

Comparison between the proposed methods and the other existing methods in moderate dimensional scenario: Setting 1.

with $a = 0.1$, and the proposed BL-SIM method, as well as the SIM-LASSO method proposed by [28], the SMAVE method proposed by [24], and the MAVE method coupled with the Bridge penalty, proposed by [25]. The last method is denoted as SIM-Bridge hereafter. For SIM-LASSO, the tuning parameter is chosen by 10-fold cross-validation, and for SMAVE and SIM-Bridge, the tuning parameter is selected based on BIC, as suggested in the original papers. Moreover, all of these three methods only use V2. In this example, we let p be moderate and vary it from 50 to 70. 100 data sets of size 100 are simulated from the following settings:

Setting 1: Model 1, COR2, and $p = 50$;
Setting 2: Model 2, COR3, and $p = 70$;
Setting 3: Model 3, COR1, and $p = 50$.

Note that Model 3 is the most difficult one, thus its dimensionality is set to 50. Under each setting, let $q = 8$, and $\theta_0 = (2, -1, 1, -0.5, 0, -1.5, 1.0, -0.3, 1.2, \cdots, 0)^T$. In this example, logGIC2 is used to choose λ. The comparison results are given in Tables 3 – 5

For both Setting 1 and Setting 2, the BS-SIM estimators outperform the rest in terms of both estimation accuracy and selection consistency. They are

Model 2, COR3, $p = 70$				
Method	Angle	FPR	TPR	Time
BS-SIM-V1	**2.250** (0.959)	**0.012**	0.999	146.390
BS-SIM-V2	2.429 (0.951)	0.025	0.999	169.485
BL-SIM-V1	7.569 (2.160)	0.694	0.994	519.186
BL-SIM-V2	5.060 (1.673)	0.728	**1.000**	494.885
SIM-LASSO-V2	6.602 (1.920)	0.684	0.993	212.528
SMAVE-V2	9.275 (4.629)	0.784	0.995	**65.740**
SIM-Bridge-V2	6.775 (4.114)	0.094	0.906	166.558

TABLE 4

Comparison between the proposed methods and the other existing methods in moderate dimensional scenario: Setting 2.

Model 3, COR1, $p = 50$				
Method	Angle	FPR	TPR	Time
BS-SIM-V1	10.003 (21.004)	0.147	0.956	466.565
BS-SIM-V2	**9.346** (19.750)	**0.142**	0.965	**218.957**
BL-SIM-V1	22.328 (27.810)	0.644	**0.979**	1037.898
BL-SIM-V2	35.757 (29.855)	0.705	**0.979**	413.221

TABLE 5

Comparison between the proposed methods and the other existing methods in moderate dimensional scenario: Setting 3.

followed by the SIM-Bridge estimator in terms of selection performance. The other three methods do not produce satisfactory performance on variable selection, as they tend to result in overfitted models. In the computational efficiency aspect, the proposed BS-SIM method is also among the best. For Setting 3, the quadratic link function is used. Since X_i's are generated from a multivariate normal distribution, they concentrate around 0. However, the MAVE based methods rely on local linear expansion, thus they do not perform well around the origin, and break down for this quadratic link function. Hence, only the results from the proposed methods are presented for this setting. It can be observed that the proposed BS-SIM method exhibits acceptable performance in each aspect, and considerably outperforms the proposed BL-SIM method. Lastly, it is also worth pointing out that satisfactory performance can be maintained for the proposed methods under other combinations of model setting and correlation structure.

Example 4. This example demonstrates the performance of the proposed estimator for large p. In this example, two choices of the dimension, $p = 200$ and $p = 400$, are examined. The other settings are $q = 10$, $n = 100$ and $\theta_0 = (2, -1, 0.5, 1, -1.5, 1.2, -0.8, 0.6, 1, -1, 0, 0, \cdots, 0)^T$. For $p = 200$, the results under all of the three aforementioned correlation structures are exhibited; for $p = 400$, the proposed method cannot produce acceptable results when there exists correlation among the predictors. Nevertheless, with more data points, the proposed BS-SIM method can still handle this high dimensional scenario with correlation among the predictors. However, we exclusively focus on COR1 and $n = 100$ for $p = 400$ here. The proposed BL-SIM method suffers greatly from overselection and is too time-consuming in the large p scenario, and SIM-LASSO

p	Model	COR	Method	Angle	FPR	TPR	Time
200	1	1	BS-SIM-V1	**5.355** (5.188)	**0.001**	**0.972**	**490.5**
			BS-SIM-V2	7.086 (8.536)	0.004	0.945	858.1
			SIM-Bridge-V2	30.585 (12.085)	0.292	0.662	2216.0
	1	2	BS-SIM-V1	**8.696** (9.836)	**0.007**	**0.919**	**870.0**
			BS-SIM-V2	10.552 (11.822)	0.021	0.894	894.8
			SIM-Bridge-V2	35.201 (11.743)	0.317	0.585	2196.0
	1	3	BS-SIM-V1	16.904 (16.423)	**0.013**	0.792	**498.6**
			BS-SIM-V2	**15.974** (17.906)	0.024	**0.808**	707.4
			SIM-Bridge-V2	47.550 (11.925)	0.438	0.381	2222.0
	2	1	BS-SIM-V2	**2.124** (0.644)	0.137	**1.000**	**1823.6**
			SIM-Bridge	3.617(2.258)	**0.041**	0.991	1841.0
	2	2	BS-SIM-V2	**2.231** (0.680)	0.039	**1.000**	**1510.9**
			SIM-Bridge-V2	4.365 (3.029)	**0.034**	0.984	2262.0
	2	3	BS-SIM-V2	**2.724** (1.497)	**0.057**	**0.999**	**1786.1**
			SIM-Bridge-V2	12.435 (9.140)	0.227	0.898	2415.0
400	1	1	BS-SIM-V1	17.533 (16.648)	0.060	0.775	2296.7
			BS-SIM-V2	**12.837** (15.665)	**0.035**	**0.855**	**1991.8**
	2	1	BS-SIM-V2	2.508 (2.258)	0.213	1.000	6519.0

TABLE 6

Performance of BS-SIM with $a = 0.1$ under several settings in high dimensional scenario.

and SMAVE break down in this example. Therefore, only the results from the proposed BS-SIM method and SIM-Bridge are presented. Since V1 poses no restriction on the magnitude of ϕ, the estimation with V1 becomes noticeably more unstable, and slower for some models, as p increases. Therefore, it is recommended to use V2 when p is large. Based on our simulation studies, V1 and V2 lead to comparable results under Model 1; whereas for Model 2, V2 is much more preferable. As for the choice of k_n, it is recommended to use $k_n^3 = \log p \log n$.

Table 6 shows the results on the four metrics. In terms of estimation accuracy and selection consistency for Model 1 and $p = 200$, the proposed BS-SIM method yields reasonably accurate estimates, while SIM-Bridge does not perform well under all of the three correlation structures. For Model 2 and $p = 200$, comparable results on selection consistency are obtained. However, the proposed BS-SIM method produces more accurate estimate than SIM-Bridge, especially under COR3. When $p = 400$, SIM-Bridge fails, while the proposed BS-SIM method can still yield satisfactory results. In terms of computational capacity, for the proposed BS-SIM method, it takes about 20 minutes on average to complete one run for $p = 200$, and takes less than two hours for $p = 400$. Considering that this amount of time encompasses the search for the optimal λ on a dense grid, this computational efficiency is still acceptable. Moreover, the proposed BS-SIM method is noticeably more efficient than SIM-Bridge in this example.

Example 5. This example focuses on the impact of the Irrepresentable Conditions. In this example, let $n = 200$, $p = 30$, $q = 6$ $N = 5$ and $\theta_0 = (2.0, -1.0, 0.5, 1.0, 0.3, -0.7, 0, \cdots, 0)^T$, and we exclusively focus on Model 1. It is clear that, for a given combination of design matrix X, link function f and true index θ_0, the Irrepresentable Conditions depend on the choice of a and the Identifiability Constraint used. The following sequence of a, $a =$

$(0.05, 0.1, 0.3, 0.5, 1.0, 2.0, 5.0)$, as well as $a = \infty$, are examined, and V1 and V2 are applied.

The simulation scheme is as follows. A covariance matrix Σ is first generated from Wishart(p,p), as done in [29]. Then we generate a sample of 100 observations of X from N(0,Σ), and standardize them. 100 normalized designs are generated in this way. Next, for each generated design, we run the following simulation 100 times. During each simulation, n copies of ε_i are sampled from $N(0, 0.3^2)$, and y_i's are calculated according to Model 1. Subsequently, the proposed method with the various choices of a specified above are applied, and the percentage of times that the applied method can identify the true model along the solution path is recorded.

Since it is difficult to quantify the Irrepresentable Conditions for BS-SIM, we compute

$$\bar{\eta}_\infty \quad = \quad 1 - \|\bar{C}_0(21)\bar{C}_0^{-1}(11)\text{sign}(\phi_0(1))\|_\infty,$$

associated with the Irrepresentable Condition for BL-SIM for each design, instead. The sign of $\bar{\eta}_\infty$ indicates whether the Irrepresentable Condition for BL-SIM holds. That is, if $\bar{\eta}_\infty > 0$, the Irrepresentable Condition for BL-SIM holds; otherwise, it fails to hold. Considering the fact that the Irrepresentable Conditions for BS-SIM are more relaxed than that for BL-SIM, $\bar{\eta}_\infty$ also implies how strongly the Irrepresentable Conditions for BS-SIM satisfy or fail, to some extent. $\bar{\eta}_\infty$ is computed for each generated design according to each Identifiability Constraint. The summary can be found in Section 5 of the Supplementary Material [32].

We first look at how the magnitude of $\bar{\eta}_\infty$ affects the performance of the proposed BL-SIM method in selecting the true model. On the two top graphs in Figure 1, the percentage of times that the true model can be identified by the proposed BL-SIM method is plotted against the corresponding $\bar{\eta}_\infty$, for the two Identifiability Constraints separately. It can be observed that the percentage increases as $\bar{\eta}_\infty$ increases, for both Identifiability Constraints. The increase is the sharpest around 0, as expected. On the two bottom graphs in Figure 1, the percentage of times of achieving selection consistency for the proposed BS-SIM method with $a = 2$ is plotted against $\bar{\eta}_\infty$, for the two Identifiability Constraints separately. It is obvious that the percentage for BS-SIM is larger than that for BL-SIM at any $\bar{\eta}_\infty$ for both constraints. It is consistent with our expectation that BS-SIM with finite a should perform better in terms of variable selection than BL-SIM.

Next, we examine how a affects the proposed method in terms of selection consistency in more detail. The average percentages of times that the true model can be selected with various choices of a are shown in Table 7. In theory, the Irrepresentable Conditions become more restrictive when a increases. Thus, it is expected that it is less likely to choose the true model when a increases. However, as indicated in Table 7, when a gets larger, the percentage of runs that the true model can be identified increases slightly first, then decreases; and when a continues to increase, the percentage for the BS-SIM estimator approaches

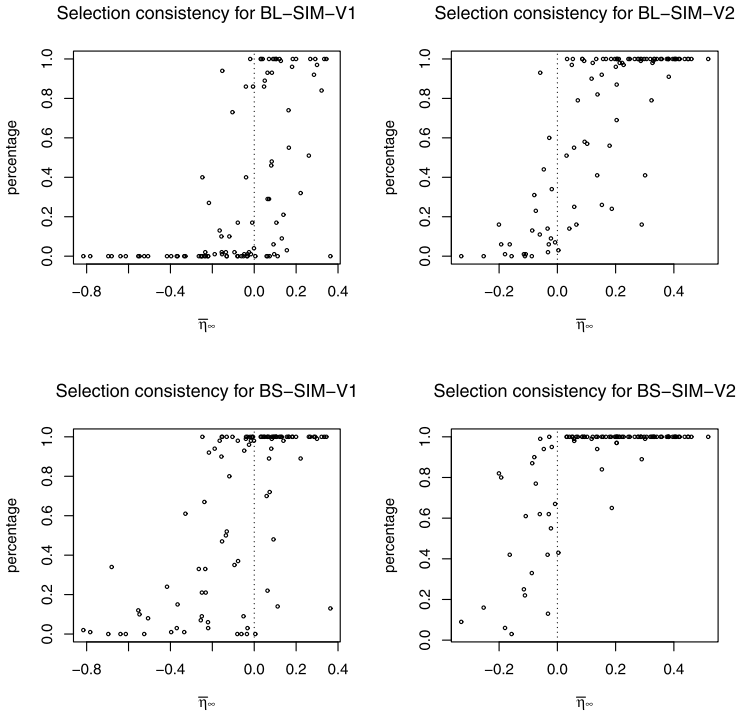

FIG 1. *The percentages that the proposed BL-SIM method and the proposed BS-SIM method with $a = 2$ select the true model versus $\bar{\eta}_\infty$ for both Identifiability Constraints.*

a	0.05	0.10	0.30	0.50	1.00	2.00	5.00	∞
V1	0.9995	1.0000	1.0000	0.9924	0.8741	0.6548	0.4665	0.3382
V2	0.9990	0.9999	0.9997	0.9956	0.9562	0.8786	0.7850	0.6909

TABLE 7
Average percentages of times that the true model can be selected with various choices of a.

that for the BL-SIM estimator. This particular pattern for the performance of BS-SIM implies that for extremely small a, it is computationally slightly more difficult to find a consistent estimator, although the Irrepresentable Conditions are relaxed. These observations on the impact of a are in line with those stated in [9].

The results in Table 7 also cast light on the role that the Identifiability Constraint plays. In most cases shown in Table 7, using V2 leads to a higher chance of recovering the true model. The difference of the chances becomes larger as a increases. This observation is consistent with the observation on the relative magnitude on $\bar{\eta}_\infty$. Among the 100 designs generated above, 92% of them have larger $\bar{\eta}_\infty$ for V2. It is probably due to the fact that the Irrepresentable Conditions for V2 contains more information than those for V1.

5. Real data application

We then apply the proposed BS-SIM method to the Skin Cutaneous Melanoma data downloaded from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/findArchives.htm). On the clinical data, there are in total 433 patients. Their demographic information, tumor status, vital status and survival time are recorded. On a separate set of files, the expression levels of 181 proteins are measured for 207 patients using the M.D. Anderson Reverse Phase Protein Array Core platform (http://www.mdanderson.org/education-and-research/resources-for-professionals/scientific-resources/core-facilities-and-services/functional-proteomics-rppa-core/index.html). The goal here is to study how the expression levels of the measured proteins influence the survival time of the patients. That means, we only retain those patients that failed to survive and had protein expression level measured for further analysis. After pre-processing, we have 94 patients, and the expression levels of 181 proteins. The expression levels are subsequently standardized and used as the predictors. The survival time is taken logarithm, and treated as the response. We apply the proposed BS-SIM method with $a = 0.1$ and 2 interior knots. Since we speculate there exists a relatively large number of relevant proteins, the GIC criterion introduced at the end of Section 2.3 with $k_n = log(n)loglog(p)$ is used to choose the tuning parameter λ. The behaviour of the logGIC criterion also to some extent confirms that the number of relevant proteins is relatively large, as it fails to effectively yield a reasonable model.

Based on the combination mentioned in the last paragraph, we are able to select 30 proteins, which are P21-R-V, 4E-BP1-pT37-T46-R-V, ACC1-R-E, Beclin-G-C, Dvl3-R-V, Notch1-R-V, p27-pT157-R-C, p53-R-E, Paxillin-R-C, PEA15-R-V, PTEN-R-V, Smad1-R-V, Smad4-M-V, Src-pY527-R-V, Syk-M-V, Tuberin-R-E, YB-1-pS102-R-V, FoxM1-R-V, MYH11-R-V, RBM15-R-V, Rictor-R-C, SCD1-M-V, TAZ-R-V, TSC1-R-C, Tuberin-pT1462-R-V, VHL-M-C, 53BP1-R-E, c-Jun-pS73-R-V, Caveolin-1-R-V and Rb-pS807-S811-R-V. The final fitted regression function is plotted against the estimated index in Figure 2.

Out of these detected proteins, the irregular expression of the p21, p27, p53, PTEN, TAZ, Notch1, Caveolin, 53BP1, TSC1, Rb and Tuberin proteins have been shown to be related to the survival or occurence of the Skin Cutaneous Melanoma [10, 15, 17, 8, 7]. This partially demonstrates the effectiveness of the proposed method in selecting the relevant variables.

6. Conclusion

In this article, we propose a regularization based approach for variable selection in the single index model, named BS-SIM. It can achieve simultanoues and efficient parameter estimation and variable selection for low to high dimensionality. A coordinate descent algorithm is outlined to implement the BS-SIM method. The algorithm is implemented in R, and the associated codes are available free online at http://www.stat.purdue.edu/~cheng70/code.html. Extensive sim-

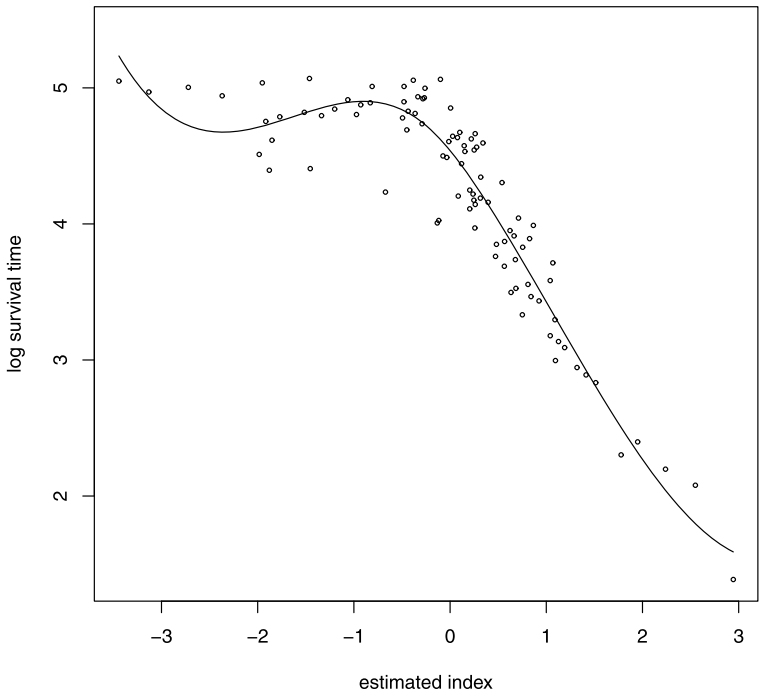

FIG 2. *The plot of the fitted regression function and the observed log survival time vesus the estimated index for the Skin Cutaneous Melanoma data.*

ulation studies are carried out to validate the proposed method and the developed algorithm. Furthermore, we show the conditions under which the BS-SIM method can consistently estimate the true index and select the true variables. These conditions generalize the conditions developed under the linear model, and are novel for the single index model.

In Section 2.3, we briefly discuss the problem of choosing the tuning parameter under different settings, and the breakdown of BIC-type method under the violation of the sparsity condition. A systematic study on the tuning parameter selection for the linear model and the single index model with a finite sample would be an interesting future research topic. Furthermore, in this work, the location and the number of knots for the cubic B-spline functions are determined by rule of thumb. There are more sophisticated methods for choosing the knots in the literature, for instance, see [13] and [30]. How to develop a novel knots placement method for BS-SIM is another interesting future research topic.

Acknowledgments

The authors would like to thank the editor, the associate editor and the referees for their insightful suggestions and comments. This research is supported by National Science Foundation under reference NSF-DMS-1107047.

Supplementary Material

Supplementary Material to "BS-SIM: An Effective Variable Selection Method for High-dimensional Single Index Model"
(doi: 10.1214/17-EJS1329SUPP; .pdf). The supplementary material contains the technical proofs and additional simulation results.

References

[1] Candes, E. J. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, Vol. 35, 2313–2404. MR2382647

[2] de Boor, C. (2001). *A Practical Guide to Splines*. Springer-Verlag, New York. MR1900298

[3] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360. MR1946581

[4] Fan, Y. and Tang, C.Y. (2013). Tuning parameter selection in high dimensional penalized likelihood *Journal of the Royal Statistical Society, Series B*, 75, Part 3, 531–552. MR3065478

[5] Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84, 986-995. MR1134488

[6] Kong E. and Xia Y.C. (2007). Variable selction for the single-index model. *Biometrika*, 94: 217–229. MR2367831

[7] Liu, Z.J., Xiao, M., Balint, K., Smalley, K.S., Brafford, P., Qiu,R., Pinnix, C.C., Li, X., and Herlyn, M. (2006). Notch1 signaling promotes primary melanoma progression by activating mitogen-activated protein kinase/phosphatidylinositol 3-kinase-Akt pathways and up-regulating N-cadherin expression. *Cancer Research*, 66, 4182–4190.

[8] Lu, M., Breyssens, H., Salter, V., Zhong, S., Hu, Y., Baer, C., Ratnayaka, I., Sullivan, A., Brown, N.R., Endicott, J., Knapp, S., Kessler, B.M., Middleton, M.R., Siebold, C., Jones, E.Y., Sviderskaya, E.V., Cebon, J., John, T,, Caballero, O.L., Goding, C.R., and Lu, X. (2013). Restoring p53 Function in Human Melanoma Cells by Inhibiting MDM2 and Cyclin B1/CDK1-Phosphorylated Nuclear iASPP. *Cancer cell*, 23(5):618-33.

[9] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, Vol. 37, No. 6A, 3498-3528. MR2549567

[10] Ming, M. and He, Y.Y. (2009). PTEN: new insights into its regulation and function in skin cancer. *Journal of Investigative Dermatology*, 129, 2109–2112.

[11] Naik, P.A. and TSAI, C.-L. (2001). Single-index model selections. *Biometrika*, 88, 821–32. MR1859412

[12] Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer-Verlag, New York. MR2244940

[13] Osborne, M.R., Presnell, B., and Turlach B.A. (1998). Knot selection for regression splines via the Lasso. *Computing Science and Statistics*, 30, 44–49.

[14] Peng, H. and Huang, T. (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, 141, 1362–1379. MR2747907

[15] Piccolo, S., Cordenonsi, M., and Dupont, S. (2013). Molecular pathways: YAP and TAZ take the center stage in organ growth and tumorigenesis. *Clinical Cancer Research*, 19(18):4925-30.

[16] Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57, 1403-1430. MR1035117

[17] Roesch, A., Becker, B., Meyer, S., Hafner, C., Wild, P.J., Landthaler, M., and Vogt, T. (2005). Overexpression and hyperphosphorylation of retinoblastoma protein in the progression of malignant melanoma. *Modern Pathology*, 18, 565–572.

[18] Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statistica Sinica*, 7, 221-264. MR1466682

[19] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6 (2), 461–464. MR0468014

[20] Tibshirani, R.J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58, 267-288. MR1379242

[21] van de Geer, S. (2000). *Empirical Processes in M-Estimation.* Cambridge University Press.

[22] Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters *Journal of the Royal Statistical Society, Series B*, 71, Part 3, 671–683. MR2749913

[23] Wang, L. and Yang, L. (2009). Spline estimation of single-index models. *Statistica Sinica*, 19, 765-783. MR2514187

[24] Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse mave. *Computational Statistics and Data Analysis*, 52, 4512–4520. MR2432477

[25] Wang, T., Xu, P., and Zhu, L. (2013). Penalized minimum average variance estimation. *Statistica Sinica*, 23, 543-569. MR3086646

[26] Xia, Y. (2006). Asymptotic distribution for two estimators of the single-index model *Econometric Theory*, 22, 1112–1137. MR2328530

[27] Xia, Y., Tong, H., Li, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space (with discussion). *Journal of the Royal Statistical Society, Series B*, 64, 363-410. MR1924297

[28] Zeng, P., He, T. and Zhu, Y. (2012). A Lasso-Type Approach for Estimation and variable Selection in Single Index Models. *Journal of Computational and Graphical Statistics*, 21, 92-109. MR2913358

[29] Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7, 2541-2563. MR2274449

[30] Zhou, S. and Shen, X. (2001). Spatially Adaptive Regression Splines and Accurate Knot Selection Schemes. *Journal of the American Statistical Association*, Vol. 96, No. 453, 247-259. MR1952735

[31] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429. MR2279469

[32] Cheng, L., Zeng, P. and Zhu, Y. (2017). Supplementary Material to "BS-SIM: An Effective Variable Selection Method for High-dimensional Single Index Model". DOI: 10.1214/17-EJS1329SUPP.