Journal of Multivariate Analysis 139 (2015) 266–282

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

High dimensional single index models

Peter Radchenko

University of Southern California, Bridge Hall 401, Los Angeles, CA, 90089, USA

ARTICLE INFO

Article history:
Received 29 April 2014
Available online 19 February 2015

AMS subject classification:
62J02
62J07
62F12

Keywords:
Nonlinear regression
High-dimensional data
Single index model
Sparsity
Regularization

ABSTRACT

This paper addresses the problem of fitting nonlinear regression models in high-dimensional situations, where the number of predictors, p, is large relative to the number of observations, n. Most of the research in this area has been conducted under the assumption that the regression function has a simple additive structure. This paper focuses instead on single index models, which are becoming increasingly popular in many scientific fields including biostatistics, economics and financial econometrics. Novel methodology is presented for estimating high-dimensional single index models and simultaneously performing variable selection. A computationally efficient algorithm is provided for constructing a solution path. Asymptotic theory is developed for the proposed estimates of the regression function and the index coefficients in the high-dimensional setting. An investigation of the empirical performance on both simulated and real data demonstrates strong performance of the proposed approach.

1. Introduction

Fields ranging from image processing and data compression to computational biology, climatology, economics and finance nowadays share the common feature of trying to extract information from vast noisy data sets. Such large scale problems may often be formulated under the framework of high-dimensional statistical regression, where the number of explanatory variables, p, is large relative to the number of observations, n. High-dimensional nonlinear regression is a research area that has recently generated a lot of interest with the emergence of powerful regularization methods. Due to the "curse of dimensionality" [1], most of the work on this subject has been performed under the assumption that the regression function has a simple additive structure. In other words, a typical regression model is

$$Y_i = \sum_{j=1}^{p} f_j^*(X_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Under the assumption that the number of "true" predictors is relatively small, this model has recently been extended into the high-dimensional setting. For example, the SpAM approach of Ravikumar et al. [28] fits a sparse additive model by imposing a penalty on the empirical L_2 norms of the functional components. Huang et al. [15] establish variable selection consistency for an adaptive variation on the SpAM approach using a B-spline implementation. Meier et al. [19] fit the same model as SpAM, but also incorporate a smoothness term in the penalty function, leading to interesting theoretical properties. Choi et al. [5] and Radchenko and James [26] extend this line of research to handle sparse high-dimensional models with interactions.

E-mail address: radchenk@usc.edu.

http://dx.doi.org/10.1016/j.jmva.2015.02.007

A different line of development uses models of the form

$$Y_i = f^*(X_i^T\boldsymbol{\alpha}^*) + \varepsilon_i, \quad i = 1, \dots, n. \tag{2}$$

Here X_i is a realization of a p-dimensional predictor vector, and linear combination $X_i^T\boldsymbol{\alpha}^*$ is referred to as index. Eq. (2) defines the Single Index Model [11,16,10]. Single index models generalize linear regression by replacing the linear predictor with a semi-parametric component. Due to their flexibility and interpretability of the coefficients single index models are becoming increasingly popular in many scientific fields. Note that they are capable of modeling interactions among predictors, and thus serve as a useful alternative to additive models.

Single index models have been very popular in relatively low and moderate dimensional situations with a manageable number of predictors. The corresponding existing methods typically do not perform well in high-dimensional situations if applied directly. Limited research, however, has been conducted on extending these methods to the high-dimensional setting. This is mainly due to the non-convexity of the sum of squares with respect to the index coefficients; an issue that makes it very hard to either develop an efficient estimation algorithm or establish theoretical properties of the estimator. Wang and Yin [34] present an approach, SMAVE, which produces sparse index coefficient estimators by introducing L_1 regularization into the MAVE method of Xia et al. [36]. However, SMAVE cannot be implemented for $p > n$, and no high-dimensional asymptotic results have been established for it. Peng and Huang [24] estimate the single index model by minimizing a penalized least squares criterion, thus performing automatic variable selection. However, they focus on the case of n being larger than p and consider only fixed p asymptotics. Also, it is argued in Section 2 that such penalization may be problematic in high-dimensional situations due to the non-convexity of the sum of squares function.

Strong interest has been generated recently by the variable selection problem under the sufficient dimension reduction framework [6]. For example, Ni et al. [23] introduce L_1 regularization to sliced inverse regression; Zhou and He [39] use L_1 regularization together with thresholding for variable filtering; Li and Yin [18] implement a sliced inverse regression approach with L_2 and L_1 regularization; Bondell and Li [3] generalize the penalization idea to a family of inverse regression estimators; Zhu and Zhu [42] investigate variable selection for single-index models with a diverging number of predictors by using inverse regression; Wu and Li [35] establish asymptotic properties for a family of inverse regression estimators in the case where p is allowed to diverge; Wang et al. [32] analyze non-convex penalized estimation in high-dimensional models with single-index structure; Yu et al. [38] use the Dantzig selector approach, together with sliced inverse regression, to perform dimension reduction and predictor selection in semi-parametric models. The dimension reduction approaches are applicable to the estimation of index coefficients in single index models. However, they do not directly estimate the regression function, and are implemented under an additional linearity condition on the distribution of the predictors.

This paper considers a different approach and makes the following key contributions:

1. A new L_1 regularization method is introduced for efficiently estimating all components of the single index model and performing variable selection simultaneously. The method is designed for the high-dimensional setting, which includes situations where p is larger than n. The level of regularization is controlled by a tuning parameter, and an algorithm is presented for constructing a solution path with respect to this parameter in a computationally efficient manner.
2. Asymptotic theory is developed for the proposed estimates of the index coefficients and the regression function in the high-dimensional setting. In particular, under some assumptions, a polynomial rate of convergence for all estimators is established even in situations where p grows faster than n.

The organization of the paper is as follows. Section 2 introduces the new methodology, provides motivation for it and discusses the intuition behind it. A computationally efficient algorithm for constructing a solution path is presented in Section 3. Theoretical investigation is conducted in Section 4. Simulation and real data performance is discussed in Section 5, an extension to the Generalized Single Index Models is presented in Section 6 and concluding remarks are given in Section 7.

2. Methodology

This section presents a new approach for fitting the single index model, (2), in situations where the number of predictors, p, is large relative to the number of observations, n. As is very common in the high-dimensional statistical inference literature, it will be assumed that the true index incorporates only a small number of predictors, in other words $\boldsymbol{\alpha}^*$ is sparse. We will refer to the proposed method as HD-SIM, which stands for High Dimensional Single Index Models.

Let $\mathbf{Y}$ denote the response vector. For a given function f and vector $\boldsymbol{\alpha} \in \mathbb{R}^p$ we will define $\mathbf{f}_{\boldsymbol{\alpha}} = \big(f(X_1^T\boldsymbol{\alpha}), \dots, f(X_n^T\boldsymbol{\alpha})\big)^T$. For a given $\boldsymbol{\alpha}$, candidate functions f will be chosen from a functional class $\mathcal{F}_n(\boldsymbol{\alpha})$. We will focus on cubic B-spline functions, but the proposed methodology works for other choices of $\mathcal{F}_n(\boldsymbol{\alpha})$. When $\boldsymbol{\alpha}$ is the jth coordinate vector, we will slightly abuse the notation and replace $\mathbf{f}_{\boldsymbol{\alpha}}$ and $\mathcal{F}_n(\boldsymbol{\alpha})$ with $\mathbf{f}_j$ and $\mathcal{F}_n(j)$, respectively. This is done for notational simplicity. Thus, we define $\mathbf{f}_j = \big(f(X_{1j}), \dots, f(X_{nj})\big)^T$, where index j refers to the jth predictor. Formal definitions of functional classes $\mathcal{F}_n(\boldsymbol{\alpha})$ and $\mathcal{F}_n(j)$ will be given in Section 3.1. For the remainder of the paper $\|\cdot\|$ will refer to the usual Euclidean vector norm, while $\|\cdot\|_1$ will correspond to the L_1 vector norm.

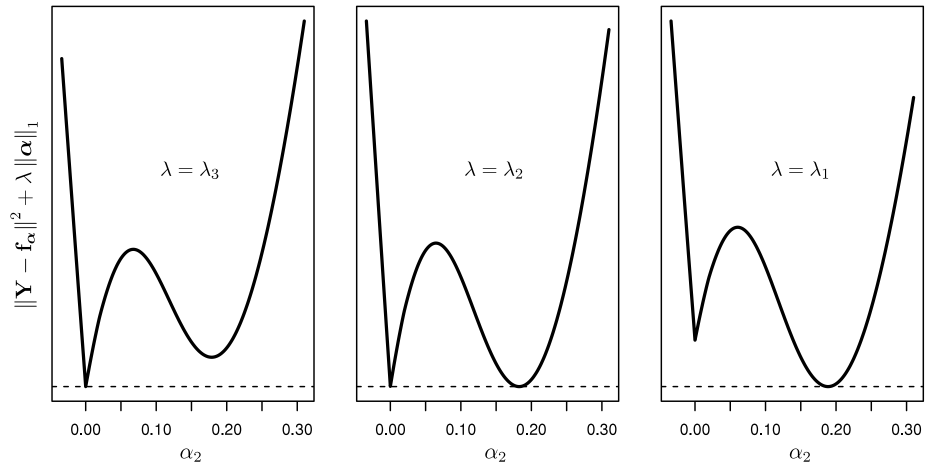

Fig. 1. Penalized least-squares criterion in Example 1 is plotted for three different values of the tuning parameter, $\lambda_3 > \lambda_2 > \lambda_1$.

There exist various approaches for estimating the index coefficients in the classical setting, such as those presented in [11,13,14,7]. We will follow the general methodology of minimizing the sum of squares [9,10,16,37],

$$\min_{\boldsymbol{\alpha},f} \|\mathbf{Y}-\mathbf{f}_{\boldsymbol{\alpha}}\|^2 ,$$

but impose additional constraints on the index coefficient vector. To achieve sparsity of the estimated coefficient vector, a natural idea is to impose an L_1 penalty on $\boldsymbol{\alpha}$:

$$\min_{\boldsymbol{\alpha},f} \|\mathbf{Y}-\mathbf{f}_{\boldsymbol{\alpha}}\|^2 + \lambda\|\boldsymbol{\alpha}\|_1 . \tag{3}$$

This is the general approach used in [24]. Of course, an additional identifiability is needed, because the index coefficients are defined up to a nonzero multiplicative factor. Such a condition will be provided when the proposed estimator is formally defined at the end of the section. However, it is shown below that even with an identifiability condition the use of criterion (3) naturally results in the problem of multiple local minima, and the corresponding global minimizer is discontinuous in the tuning parameter λ. This makes finding the solution extremely hard, especially in a high-dimensional situation. Consider the following simulated example for illustration.

Example 1. The response is generated from the model

$$Y_i = (X_{i1} + 0.7X_{i2} - 1)^2 + 0.2\varepsilon_i, \quad i = 1, \ldots, 150,$$

where X_{ij} are independent Uniform random variables on $[0, 1]$, and ε_i are i.i.d. $N(0, 1)$.

Under any reasonable identifiability condition, as λ is decreased, the solution path for problem (3) moves from the sparsest solution towards an un-regularized one. Note that the index for the sparsest solution is formed by the variable that is the best marginal predictor of the response. In the realization that we consider such predictor is $\mathbf{X}_1$, which is a typical scenario, as this predictor has the largest coefficient in the true model. Because the index coefficients are defined up to a nonzero multiplicative factor, we will fix the coefficient $\widehat{\alpha}_1$ at one, for identifiability. With $\widehat{\alpha}_1$ fixed, we will examine what happens to the coefficient $\widehat{\alpha}_2$ as λ is decreased. Fig. 1 plots the penalized criterion function, optimized over f, versus α_2. The plot is produced for three values of the tuning parameter, $\lambda_3 > \lambda_2 > \lambda_1$. For each of the values the criterion has two local minima. When λ is decreased past the value λ_2, the global minimizer, $\widehat{\alpha}_2$, jumps from zero to about 0.185. Note that the corresponding sum of squares function, $G(\alpha_2)$, displayed at the top of Fig. 2, is strictly decreasing on the domain of interest. Presence of multiple local minima for the penalized criterion is implied by fact that at some positive α_2 the derivative, G', displayed at the bottom of Fig. 2, drops below its value at zero. The following result, proved in Appendix E, makes this statement precise.

Proposition 1. *Consider a smooth univariate criterion function $G(\alpha)$. Suppose that G has a minimum at $\alpha^* \neq 0$, and α^* is the only zero of G'. Define a penalized criterion function $F_\lambda(\alpha) = G(\alpha) + \lambda|\alpha|$. If $G''(0) < 0$, then there exist positive values $\lambda_3 > \lambda_2 > \lambda_1$ and a, such that for all $\lambda \in (\lambda_1, \lambda_3)$ function F_λ has multiple local minima, one at zero, and another at some α_λ with $|\alpha_\lambda| > a$. For $\lambda \geq \lambda_2$ function F_λ has a global minimum at zero, and for $\lambda \leq \lambda_2$ function F_λ has a global minimum at some α_λ with $|\alpha_\lambda| > a$.*

Condition $G''(0) < 0$ can replaced by a weaker one, $|G'(0)| < \sup_{\{0<\alpha\alpha^*<|\alpha^*|^2\}} |G'(\alpha)|$.

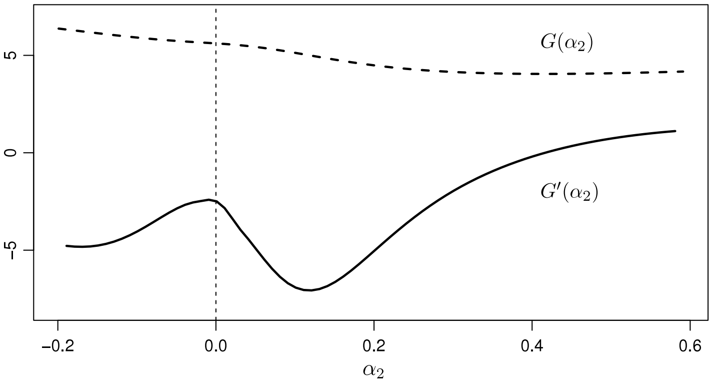

Fig. 2. Sum of squares function $G(\alpha_2) = \min_f \left\{\|\mathbf{Y} - \mathbf{f}_{\boldsymbol{\alpha}}\|^2\,,\ \boldsymbol{\alpha} = (1, \alpha_2)^T\right\}$ and its derivative for Example 1.

Thus, due to the non-convexity of the sum of squares function, the phenomenon illustrated in Fig. 1 is an inherent property of the penalized least-squares approach in single index models. This property is highly unsatisfactory from a computational point of view, because finding the solution to optimization problem (3) becomes extremely hard, especially in the high-dimensional setting. It also makes constructing a solution path computationally unfeasible, because even for a small change in the tuning parameter a global search for the minimizer of the penalized criterion function would need to be conducted. Finally, it results in discontinuity of the global solution with respect to the tuning parameter, raising concerns about the stability of the corresponding estimator.

Instead, the proposed HD-SIM approach uses an L_1 equality constraint:

$$\min_{\boldsymbol{\alpha},f} \|\mathbf{Y} - \mathbf{f}_{\boldsymbol{\alpha}}\|^2 \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 = t. \tag{4}$$

Interestingly, replacing an L_1 penalty with an L_1 constraint changes the set of solutions to the optimization problem. This phenomenon is caused by the non-convexity of the sum of squares function with respect to $\boldsymbol{\alpha}$. Consider the setting of Proposition 1. The fact that $G(\alpha)$ is decreasing for $\alpha < \alpha^*$ and increasing for $\alpha > \alpha^*$ implies that the constrained optimization problem, $\min G(\alpha)$ s.t. $|\alpha| = t$, has a unique solution given by $t\,\text{sign}(\alpha^*)$ for each t in $[0, |\alpha^*|]$. Consequently, as t is increased, the solution is *continuously* “un-shrunk” from zero to the unconstrained minimizer of G. A similar continuity property holds for the constrained solution in Example 1. It is again implied by a monotonicity argument applied to the sum of squares function, G, displayed in Fig. 2.

More generally, as Example 1 and Proposition 1 illustrate, non-convexity of the sum of squares with respect to the index coefficients tends to cause jumps in the L_1 norm of the solution to the penalized least-squares optimization problem. Such jumps are ruled out for the constraint approach, as the L_1 norm of the solution, by definition, is a continuous function (identity) of the tuning parameter. The following general result provides high-level conditions for continuity of the path of solutions to problem (4). The proof of this result is provided in the Appendix.

Proposition 2. *Let f be a fixed continuous function. Assume that for some positive T_1 and T_2, and for each $t \in [T_1, T_2]$, optimization problem (4) has a unique global solution. Then the path of solutions to this problem is continuous on $[T_1, T_2]$.*

Proposition 2 only imposes conditions on the global solutions to the optimization problem. Non-convexity of the sum of squares function does not cause discontinuity in the solution path as long as the global constrained minimizer of the sum of squares is unique. This is quite different from the situation for the penalized approach.

Note that we need to impose an additional identifiability condition in optimization problem (4). Otherwise, provided the class of candidate functions f is sufficiently wide, the $\boldsymbol{\alpha}$-solution would simply equal a rescaled version of the un-regularized minimizer of the sum of squares. Our approach will be to start the solution path at an index coefficient vector of highest sparsity, with only one nonzero coefficient, then fix this coefficient for the remainder of the solution path. More specifically, HD-SIM estimator $(\widehat{\boldsymbol{\alpha}}, \widehat{f})$ solves the following optimization problem,

$$\begin{aligned} &\min_{\boldsymbol{\alpha}\in\mathbb{R}^p, f\in\mathcal{F}_n(\boldsymbol{\alpha})} \|\mathbf{Y} - \mathbf{f}_{\boldsymbol{\alpha}}\|^2 \quad \text{s.t.} \ \|\boldsymbol{\alpha}\|_1 = t \quad \text{for } t \ge 1, \\ &\alpha_{\widehat{m}} = 1 \text{ for } \widehat{m} = \arg\min_{j\le p} \min_{f\in\mathcal{F}_n(j)} \left\|\mathbf{Y} - \mathbf{f}_j\right\|^2. \end{aligned} \tag{5}$$

The notation used in the above display was introduced in the beginning of the section. Functional classes $\mathcal{F}_n(\boldsymbol{\alpha})$ and $\mathcal{F}_n(j)$ will be formally defined in Section 3. Note that the starting point of the HD-SIM solution path, which corresponds to the value $t = 1$ of the tuning parameter, is the same as it would be if $\|\boldsymbol{\alpha}\| = 1$, rather than $\alpha_{\widehat{m}} = 1$, were used as the identifiability condition. HD-SIM solution path starts at the index formed by the variable that is the best marginal predictor of the response, fixes the respective index coefficient at one, then gradually un-shrinks the remaining coefficients and introduces new variables into the index. An algorithm for constructing a path of solutions over a dense tuning parameter grid is given in Section 3. In all of the numerical examples that follow, the optimal value of the tuning parameter is selected through cross validation.

3. Path fitting algorithm

Suppose that for each $t \in [1, T]$ optimization problem (5) has a unique global solution. If the class $\cup_{\|\boldsymbol{\alpha}\| \le T} \mathcal{F}_n(\boldsymbol{\alpha})$ consists of continuous functions and is continuously parameterized by a bounded subset of a Euclidean space, then the solution path is continuous on $[1, T]$. This follows directly from the proof of Proposition 2 after accounting for the optimization over the regression function, f. The path algorithm proposed in this section computes solutions over a very dense grid of t's, gradually increasing the tuning parameter from the starting value $t = 1$. The last computed solution is used as a warm start in the search for the next one. Under the setting described above, this search only needs to be conducted locally. Thus, we can take advantage of the local quadratic approximation to the sum of squares function. Note that the solution at $t = 1$ corresponds to the index formed by $X_{\widehat{m}}$, where $\widehat{m}$ is defined in (5). Variable $X_{\widehat{m}}$ is the best marginal predictor of the response, and can be easily identified by solving a convex optimization problem. Computational details are provided in Section 3.1.

As described in, for example, [12], existing algorithms for minimizing $\|\mathbf{Y} - \mathbf{f}_{\boldsymbol{\alpha}}\|^2$ in low-dimensional scenarios typically iterate between estimating the functional link (f update) and estimating the index coefficient vector ($\boldsymbol{\alpha}$ update). We will also follow this approach for each given value of the tuning parameter. The success of such algorithms depends on the initialization, i.e. convergence can be achieved under reasonable regularity conditions if the starting point is close to the global minimizer. Note that this is indeed the case under the setting described in the previous paragraph, provided the grid of tuning parameter values is sufficiently dense.

3.1. f update

The f update is done for a fixed value of $\boldsymbol{\alpha}$. While there are many possible ways to estimate f, we will focus on the B-spline approach following, for example, [37,33]. For a given $\boldsymbol{\alpha}$, candidate functions f will be chosen from the space of cubic splines, corresponding to the partitioning of $[\min_i X_i^T\boldsymbol{\alpha}, \max_i X_i^T\boldsymbol{\alpha}]$ into $d_n - 3$ intervals of equal length. This space will be denoted by $\mathcal{F}_n(\boldsymbol{\alpha})$. There exists [29] a normalized B-spline basis $\{b_1, \dots, b_{d_n}\}$, such that every $f \in \mathcal{F}_n(\boldsymbol{\alpha})$ can be represented as

$$f(s) = \sum_{k=1}^{d_n} \beta_k b_k(s).$$

Let $\mathsf{B}_{\boldsymbol{\alpha}}$ denote the matrix whose ith row is given by $(b_1(X_i^T\boldsymbol{\alpha}), \dots, b_{d_n}(X_i^T\boldsymbol{\alpha}))$. For every candidate function $f \in \mathcal{F}_n(\boldsymbol{\alpha})$ we can write $\mathbf{f}_{\boldsymbol{\alpha}} = \mathsf{B}_{\boldsymbol{\alpha}}\boldsymbol{\beta}$. Consequently, the f update becomes a simple optimization problem,

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathsf{B}_{\boldsymbol{\alpha}}\boldsymbol{\beta}\|^2,$$

which can be solved using ordinary least-squares. We will slightly abuse the notation and write $\mathcal{F}_n(j)$ for the space of cubic splines defined analogously on $[\min_i X_{ij}, \max_i X_{ij}]$. Ordinary least-squares can be applied to find the solution to $\min_{f \in \mathcal{F}_n(j)} \|\mathbf{Y} - \mathbf{f}_j\|^2$ for each j, which can then be used to identify $\widehat{m}$ and, thus, the starting point on the solution path.

The theory in Section 4 provides reasonable rates of growth for the number of basis functions, d_n, when the sample size tends to infinity. In particular, Corollary 1 suggests that some over-smoothing relative to the optimal nonparametric choice $d_n \sim n^{1/5}$ results in better convergence rates for $\widehat{\boldsymbol{\alpha}}$ and $\widehat{f}$. We should also note that for irregularly distributed indexes, non-uniform placement of the B-spline knots may be advantageous.

3.2. $\boldsymbol{\alpha}$ update

Once the tuning parameter, t, moves to a new grid point, t_{new}, we increase the magnitudes of some of the nonzero coefficients from their current values, to ensure that the constraint $\|\boldsymbol{\alpha}\|_1 = t_{new}$ is satisfied. It is proposed to implement the $\boldsymbol{\alpha}$ update by using a block coordinate descent approach. The basic idea is to iteratively update the coefficients in blocks of two, e.g. (j, L), where $L \neq \widehat{m}$ is some fixed index that corresponds to a large coefficient, and j runs through the index set but does not equal either $\widehat{m}$ or L. The sum $|\alpha_j| + |\alpha_L|$ is required to remain constant throughout the iteration, which maintains the L_1 constraint, while α_j and α_L are modified in order to minimize the current approximation to the objective function.

The $\boldsymbol{\alpha}$ update is typically conducted [12] by minimizing the objective function in a small neighborhood of the current $\boldsymbol{\alpha}$, where the deviation in f is replaced by a first order approximation. After the approximation the objective function takes the form $\left\|\mathbf{R} - \mathbf{f}' * (X\Delta)\right\|^2$. Here X is the matrix of predictors, $\mathbf{R} = \mathbf{Y} - \mathbf{f}_{\boldsymbol{\alpha}}$ is the vector of current residuals, vector $\mathbf{f}'$ contains derivatives of f evaluated at the current values of the index, Δ is the change in vector α that is being considered, and operation $*$ denotes element-wise multiplication of vectors. We will write $\mathbf{X}_j$ for the jth predictor vector and denote by ∇_j the gradient of $\frac{1}{2}\|\mathbf{Y} - \mathbf{f}_{\boldsymbol{\alpha}}\|^2$ with respect to α_j. In other words, we define

$$\nabla_j = -\mathbf{R}^T \left(\mathbf{f}' * \mathbf{X}_j\right). \tag{6}$$

Suppose that either the current value of α_j is not equal to zero or the inequality $|\nabla_j| > |\nabla_L|$ is satisfied. Also suppose that after the (j, L)-block update neither α_j nor α_L flips its sign. Then, expressing α_L in terms of α_j and setting the derivative of the objective function to zero yields (Appendix F) the following formula for the change in α_j,

$$\Delta_j = \frac{-\left(\nabla_j - S_{jL}\nabla_L\right)}{\left\|\mathbf{f}' * \left(\mathbf{X}_j - S_{jL}\mathbf{X}_L\right)\right\|^2}, \quad \text{with } S_{jL} = \operatorname{sign}(\alpha_L\alpha_j) - \operatorname{sign}(\alpha_L\nabla_j)1_{\{\alpha_j=0\}}. \tag{7}$$

The value of Δ_L is then determined through the constraint on $|\alpha_j| + |\alpha_L|$. A situation where a coefficient crosses zero can be handled by setting that coefficient to exactly zero and correspondingly updating the other coefficient in the block.

We can speed up the $\boldsymbol{\alpha}$ update by first iterating through the coefficients that are currently nonzero, and then checking whether any of the excluded predictors should be added to the model. The details of the $\boldsymbol{\alpha}$ update are provided in Algorithm 1 below. Symbol $\mathcal{A}$ denotes the current "active" index set, which identifies the nonzero coefficients of $\boldsymbol{\alpha}$. Because the path is constructed on a very dense grid of tuning parameter values, for most of the grid points the active set does not change. Consequently, step B of the algorithm is typically implemented until convergence only once and only involves the currently nonzero coefficients of $\boldsymbol{\alpha}$. This results in significant computational savings when p is large, especially on the sparse part of the solution path. It is also suggested that the algorithm be modified by performing the f update as a sub-step (e) in step B of Algorithm 1, rather than implementing the $\boldsymbol{\alpha}$ update iterations until convergence before updating f. Simulations have shown that this modification produces some computational gains.

Algorithm 1 $\boldsymbol{\alpha}$ update

```
A. Set $s_j = \operatorname{sign}(\alpha_j)$ for each $j$; if $s_k = 0$ for a $k$ in $\mathcal{A}$, set $s_k = -\operatorname{sign}(\nabla_k)$ using (6).
   If $|\mathcal{A}| = 1$ define $L = \arg\max_{j \in \mathcal{A}^c} |\nabla_j|$, otherwise let $L = \arg\max_{j \in \mathcal{A}\setminus\{\widehat{m}\}} |\alpha_j|$.
      Iterate step B until convergence.
B. For each $j$ in $\mathcal{A} \setminus \{\widehat{m}, L\}$ do
   (a) if either $\alpha_j \neq 0$ or each of the inequalities $s_j\nabla_j < 0$ and $|\nabla_j| > |\nabla_L|$ is satisfied, set $\alpha_j = \alpha_j + \Delta_j$, where $\Delta_j$ is defined
       in (7); otherwise set $\Delta_j = 0$
   (b) if $\alpha_j$ switched sign in (a), set $\Delta_j = \Delta_j - \alpha_j$ and $\alpha_j = 0$
   (c) set $\alpha_L = \alpha_L - \Delta_j \operatorname{sign}(s_j\alpha_L)$
   (d) if $\alpha_L$ switched its sign in (c), set $\alpha_j = \alpha_j + |\alpha_L|s_j,\ \alpha_L = 0$, and redefine $L$.
C. At this point, $|\nabla_j| = |\nabla_L|$ for all $j \in \mathcal{A} \setminus \{\widehat{m}\}$. If there exists an index $k \in \mathcal{A}^c$ for which $|\nabla_k| \geq |\nabla_L|$, augment $\mathcal{A}$ with $k$ and
   go to step A. Otherwise, stop.
```

4. Theoretical investigation

This section develops asymptotic theory for the proposed estimates of the index coefficients and the regression function in the high-dimensional setting. A rate of convergence for all estimators (Theorem 2) is established even in situations where p grows faster than n. Existing theory for high-dimensional L_1 regularized estimation cannot be used for the HD-SIM estimator of the index coefficients, because function $\|\mathbf{Y} - \mathbf{f}_{\boldsymbol{\alpha}}\|^2$ is not convex with respect to $\boldsymbol{\alpha}$. In particular, results on high-dimensional rates of convergence, e.g. [2,21], rely on the convexity of the objective function. Estimation theory for high-dimensional convex problems follows from a restricted strong convexity condition [22]. For a non-convex objective function such an assumption is not feasible. Instead, we first establish consistency of the estimator (Theorem 1) and then impose a local version of the restricted strong convexity condition. While it is not explicitly stated in the results below, the proofs of Theorems 1 and 2 also establish consistency and the rate of convergence for $\widehat{\mathbf{f}}_{\widehat{\boldsymbol{\alpha}}}$, the estimator of the regression surface represented by $\mathbf{f}^*_{\boldsymbol{\alpha}^*}$. All the results in this section are given under the assumption of the single index model, (2). Without this assumption, one could conduct the asymptotic analysis using the ideas from semiparametric M-estimation and following, for example, the approach in [17].

One of the conditions required for the results in this section is condition A7, which is given in the Appendix. It implies that the best marginal predictor of the response, $\mathbf{X}_{\widehat{m}}$, is one of the signal predictors with probability tending to one. It follows from the proof of Theorem 1 that this assumption is not needed for the consistency of $\widehat{\mathbf{f}}_{\widehat{\boldsymbol{\alpha}}}$. However, it is required for the consistency of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{f}$. Interestingly, assumption A7 can be removed from all the results in this section if we replace the identifiability constraint $\alpha_{\widehat{m}} = 1$ with $\|\boldsymbol{\alpha}\| = 1$ in the optimization problem (5). Thus, the extra condition can be viewed as the price that is paid for the computational simplicity of maintaining the trivial constraint $\alpha_{\widehat{m}} = 1$ relative to the constraint $\|\boldsymbol{\alpha}\| = 1$.

Note that the true and the estimated index coefficient vectors, $\boldsymbol{\alpha}^*$ and $\widehat{\boldsymbol{\alpha}}$, are identifiable up to nonzero multiplicative factors. For concreteness, we will assume that $\|\boldsymbol{\alpha}^*\|_1 = 1$. Whenever $X_{\widehat{m}}$ is a signal predictor, we will rescale $\widehat{\boldsymbol{\alpha}}$ to achieve $\widehat{\alpha}_{\widehat{m}} = \alpha^*_{\widehat{m}}$. All the results in this section correspond to this scaling of $\boldsymbol{\alpha}^*$ and $\widehat{\boldsymbol{\alpha}}$. Throughout this section we will treat the predictors as deterministic. This is a typical approach in the asymptotic analysis of the nonlinear least-squares [30]. Recall that d_n denotes the number of basis functions used in computing the HD-SIM estimator. Let Q_n denote the empirical probability measure associated with the true index values $X_1^T\boldsymbol{\alpha}^*, \ldots, X_n^T\boldsymbol{\alpha}^*$. Theorem 1 demonstrates that the proposed approach can consistently estimate the index coefficients and the regression function of the true model.

Theorem 1. *Suppose* $d_n \to \infty$, $d_n = o\big(n^{1/2}[\log n]^{-3/2}\big)$ *and* $p_n = o\big(\exp(n[d_n \log n]^{-2})\big)$, *as* n *tends to infinity. If regularity conditions* A1–A7 *in* Appendix A *are satisfied, then the HD-SIM tuning parameter can be chosen in such a way that*

$$\big|\big|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\big|\big| = o_p\big(1\big) \quad \text{and} \quad \|\widehat{f} - f^*\|_{L_2(\mathbb{Q}_n)} = o_p\big(1\big). \tag{8}$$

In particular, if we let d_n grow at the optimal nonparametric rate $n^{1/5}$, then all the components of the single index model are estimated consistently even if p_n increases at the rate $\exp(n^{3/5-\tau})$ for some arbitrarily small positive τ. It follows from the proof that the appropriate value of the tuning parameter, t, can be selected from an interval of size $o_p(1)$. The location of the interval, specified in the proof, depends on unknown quantities associated with the true index coefficient vector. Note that the same phenomenon occurs in the widely-used constrained formulation of the Lasso criterion in the linear regression setting, min $\|\mathbf{Y} - X\beta\|^2$, s.t. $\|\boldsymbol{\beta}\|_1 \leq s$. To achieve consistency in the high-dimensional setting, s needs to be selected from a $o_p(1)$ neighborhood of $\|\boldsymbol{\beta}^*\|_1$, the L_1 norm of the true coefficient vector. In all the empirical work in this paper, the value of the tuning parameter is chosen by cross validation.

The next theorem addresses the rate of convergence of the HD-SIM estimator. To derive this result we need to impose a regularity condition controlling the behavior of the sum of squares function near its minimum. The new assumption, B, which is given and discussed in Appendix A, is a generalization of the conditions required for the Lasso estimation theory. Let $\mathcal{A}^*$ denote the index sets for the nonzero coefficients of $\boldsymbol{\alpha}^*$, and define $s_n = |\mathcal{A}^*|$.

Theorem 2. *Let* $r_n = n^{-1/2}s_n^{1/2}\log n\sqrt{\log p_n} + n^{-1/2}\sqrt{d_n \log n} + d_n^{-2}$. *Suppose* $d_n \to \infty$, $d_n = o\big([n \log n]^{1/4}\big)$, $p_n = o\big(\exp(ns_n^{-1}d_n^{-3}[\log n]^{-2})\big)$, *and assumptions* A1–A7 *and* B *are satisfied. Then the HD-SIM tuning parameter can be chosen in such a way that*

$$\big|\big|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\big|\big| = O\big(r_n\big) \quad \text{and} \quad \|\widehat{f} - f^*\|_{L_2(\mathbb{Q}_n)} = O\big(r_n\big),$$

with probability tending to one, as n *goes to infinity.*

The following result considers some concrete growth rates for p_n, and optimizes the rate of convergence in Theorem 2 over the choice of d_n.

Corollary 1. *Suppose that assumptions* A1–A7 , B *are satisfied and* $d_n \sim n^{1/5}(\log n)^{-1/5}$.

1. *Let* $p_n = O\big(\exp(s_n^{-1}n^{1/5-\epsilon})\big)$ *for some arbitrarily small positive* ϵ. *Then, the result of* Theorem 2 *holds with* $r_n = n^{-2/5}[\log n]^{2/5}$.
2. *Let* $p_n = \exp(s_n^{-1}n^c)$ *for* $c \in [1/5, 2/5)$. *Then, the result of* Theorem 2 *holds with* $r_n = n^{-(1-c)/2}\log n$.

Note that a polynomial rate of convergence can be achieved even when p_n is growing as fast as $\exp(s_n^{-1}n^{2/5-\epsilon})$, for an arbitrarily small positive ϵ. However, it is important to point out that the HD-SIM methodology is not recommended for practical applications in ultra-high dimensional settings. The rate of convergence in these settings is slow. A more reasonable approach would be to first apply a variable screening procedure, such as those developed in [41,8]. Then, after the dimension of the problem has been reduced from ultra-high to high, one can apply the HD-SIM approach.

We now turn to the variable selection properties of the HD-SIM estimator. As discussed in Section 5, methods that use L_1 regularization are known to produce models containing a large number of noise predictors. To alleviate this problem, consider the popular approach of thresholding the initial estimator. Let $q_n = n^{-1/2}\log n\sqrt{\log p_n} + n^{-1/2}\sqrt{d_n \log n} + d_n^{-2}$. Define $\widetilde{\boldsymbol{\alpha}}$, the thresholded HD-SIM estimator of the index coefficient vector, as follows: $\widetilde{\alpha}_j = \widehat{\alpha}_j I\{|\widehat{\alpha}_j| > q_n\}, j = 1, \ldots, p_n$. Let $\widetilde{\mathcal{A}}_n$ denote the index set of the corresponding selected predictors: $\{j : 1 \leq j \leq p_n, \widetilde{\alpha}_j \neq 0\}$. The next result provides bounds for the estimation error of the thresholded HD-SIM approach and for the corresponding number of selected predictors.

Theorem 3. *Under the settings of* Theorem 2,

$$|\widetilde{\mathcal{A}}_n| = O(|\mathcal{A}^*|) \quad \text{and} \quad \big|\big|\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\big|\big| = O\big(s_n^{1/2}q_n\big),$$

with probability tending to one, as n tends to infinity.

Now consider the case where the components of the true model do not depend on n. More specifically, suppose that the number of signal predictors, $|\mathcal{A}^*|$, and the corresponding index coefficients are fixed and do not change with n. The estimation error bound in Theorem 3 implies that, with probability tending to one, the thresholded estimator has zero false negatives, while the number of false positives stays bounded. This variable selection result can be strengthened by increasing the threshold from q_n to aq_n, for a sufficiently large a. In this case, the corresponding thresholded estimator can correctly recover the index set of the relevant predictors, with probability tending to one. Using only the selected predictors, one could produce unregularized estimators of $\boldsymbol{\alpha}^*$ and f^* by applying, for example, the methods in [37] or [33]. The asymptotic normality results developed in the corresponding paper would then hold for the final estimator, allowing one to conduct inference for the components of the true model.

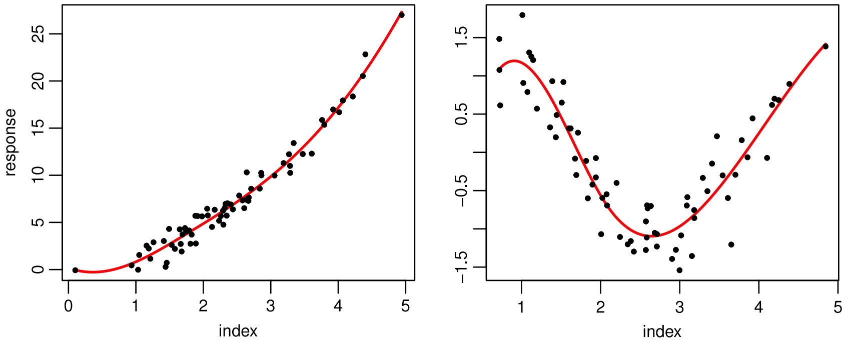

Fig. 3. Estimated regression function and observed values of the response variable are plotted against an HD-SIM estimated index for two simulation runs.

5. Empirical analysis

In this section the finite-sample performance of the proposed approach is compared with some of the strong competitors discussed in the Introduction. More specifically, we investigate how HD-SIM performs relative to SMAVE, a method of Wang and Yin [34], and three different variants of SR-SIR, which is an approach of Li and Yin [18]. All of the methods utilize some form of L_1 regularization in estimating the index coefficients and performing variable selection.

It has been observed for some time that one of the limitations of shrinkage methods, such as the Lasso, is that in situations where the true number of non-zero coefficients is small relative to p, they must choose between including a large number of irrelevant variables or else over-shrinking the coefficients. This tradeoff is caused by the fact that these methods use a single tuning parameter to control both the variable selection and the shrinkage component of the fitting procedure. A way to overcome this issue is to post-process the shrinkage estimator ([4,20,27], for example). One popular and simple approach is to compute un-regularized solutions using only the variables selected by the original shrinkage estimator. We can post-process the HD-SIM estimator along these lines. More specifically, for every given value of the tuning parameter, the unregularized least-squares problem is solved while restricting the choice of predictors to only those selected by the original estimator. The minimization is performed by using the original estimator as a warm start and then iterating the f update and the unconstrained $\boldsymbol{\alpha}$ update until convergence. The final estimator, to which we will refer as HD-SIM-pst, is selected as a point on the resulting path of post-processed HD-SIM solutions through cross validation. Both the original and the post-processed HD-SIM estimators are examined in the simulation study in Section 5.1, however, the main focus of the comparisons is on the performance of the original estimator. The HD-SIM methods use $d_n = 6$ B-spline basis functions in the estimation, which is an adequate number for estimating reasonably smooth link functions. Five-fold cross validation is performed to select the values of the tuning parameter. The SMAVE and SR-SIR approaches are implemented using the code provided by the authors of the corresponding papers.

5.1. Simulation study

We consider two types of single index models for generating the response, a quadratic model and an oscillating function model. The following two scenarios are taken from [7], however the sample size is significantly reduced, and the noise level is increased in half of the simulation runs. Also, while the distribution of the true index is kept intact, the true coefficient values are redistributed among four predictors, as opposed to two in the original simulation setting.

Simulation scenario 1. The data is generated according to a simple quadratic model,

$$Y_i = \left(X_i^T \boldsymbol{\alpha}^*\right)^2 + \sigma\varepsilon_i, \quad X_i \sim N_p(2\sqrt{17}/5, I), \; i = 1, \dots, n.$$

Simulation scenario 2. The data comes from the following oscillating function model,

$$Y_i = \sin\left(X_i^T \boldsymbol{\alpha}^* \pi/2\right) + \sigma\varepsilon_i, \quad X_i \sim N_p(2\sqrt{17}/5, I), \; i = 1, \dots, n.$$

For each scenario the number of observations and the standard deviation of the noise are varied. The number of observations is taken as either $n = 140$ or $n = 70$, while p is fixed at 100. This way both the $p < n$ and the $p > n$ situations are considered. For all of the methods 100 simulation runs are performed for each setting. For each simulation scenario, the true index coefficient vector, $\boldsymbol{\alpha}^*$, is set to $(8, 4, 2, 1, 0, \dots, 0)/\sqrt{85}$. The error terms, ε_i, are independently produced from the standard normal distribution, and the predictor vectors, X_i, are generated from a multivariate normal distribution. Fig. 3 displays the estimated regression function and observed values of the response variable plotted against the HD-SIM estimated index for two simulation runs. The left plot corresponds to Simulation scenario 1 with $n = 70$ and $\sigma = 1$, while the right plot corresponds to Simulation scenario 2 with $n = 70$ and $\sigma = 0.4$.

Table 1
Average estimation errors (with standard errors in parenthesis), average number of false positives and average number of false negatives for simulation scenarios 1 and 2.

Simulation setting	Method	Estimation error	False positives	False negatives
	SMAVE	0.093 (0.005)	2.96	0
Quadratic	SR-SIR AIC	0.249 (0.012)	1.99	0.39
$p = 100$	SR-SIR BIC	0.271 (0.010)	0.27	0.73
$n = 140$	SR-SIR RIC	0.316 (0.011)	0.05	0.94
$\sigma = 0.2$	HD-SIM	0.042 (0.001)	12.96	0
	HD-SIM-pst	0.010 (0.001)	0.24	0
	SMAVE	0.157 (0.006)	5.31	0
Quadratic	SR-SIR AIC	0.304 (0.013)	2.97	0.55
$p = 100$	SR-SIR BIC	0.304 (0.009)	0.38	0.92
$n = 140$	SR-SIR RIC	0.326 (0.009)	0.10	1.01
$\sigma = 1$	HD-SIM	0.199 (0.008)	12.15	0
	HD-SIM-pst	0.061 (0.005)	0.49	0.01
	SR-SIR AIC	0.844 (0.085)	6.06	1.08
Quadratic	SR-SIR BIC	0.602 (0.060)	1.65	1.47
$p = 100$	SR-SIR RIC	0.563 (0.049)	0.66	1.65
$n = 70$	HD-SIM	0.082 (0.013)	13.35	0.01
$\sigma = 0.2$	HD-SIM-pst	0.030 (0.010)	0.73	0.02
	SR-SIR AIC	0.835 (0.052)	6.72	1.01
Quadratic	SR-SIR BIC	0.580 (0.034)	1.42	1.57
$p = 100$	SR-SIR RIC	0.553 (0.022)	0.59	1.75
$n = 70$	HD-SIM	0.340 (0.015)	12.31	0.06
$\sigma = 1$	HD-SIM-pst	0.163 (0.012)	1.55	0.37
	SMAVE	4.188 (0.201)	26.61	1.61
Oscillating	SR-SIR AIC	5.024 (0.114)	26.65	2.15
$p = 100$	SR-SIR BIC	3.164 (0.113)	8.20	2.84
$n = 140$	SR-SIR RIC	2.399 (0.104)	3.09	3.12
$\sigma = 0.2$	HD-SIM	0.237 (0.010)	12.29	0
	HD-SIM-pst	0.098 (0.010)	1.05	0.05
	SMAVE	4.822 (0.143)	31.22	1.89
Oscillating	SR-SIR AIC	5.327 (0.098)	28.37	2.16
$p = 100$	SR-SIR BIC	3.360 (0.111)	8.48	2.94
$n = 140$	SR-SIR RIC	2.676 (0.103)	3.70	3.33
$\sigma = 0.4$	HD-SIM	0.387 (0.013)	9.07	0.17
	HD-SIM-pst	0.200 (0.012)	0.85	0.57
	SR-SIR AIC	5.195 (0.078)	23.85	2.63
Oscillating	SR-SIR BIC	3.856 (0.091)	10.24	3.11
$p = 100$	SR-SIR RIC	3.058 (0.089)	4.36	3.51
$n = 70$	HD-SIM	0.433 (0.039)	8.38	0.32
$\sigma = 0.2$	HD-SIM-pst	0.293 (0.044)	1.48	0.68
	SR-SIR AIC	5.329 (0.080)	24.34	2.67
Oscillating	SR-SIR BIC	3.924 (0.094)	10.15	3.25
$p = 100$	SR-SIR RIC	3.167 (0.078)	4.52	3.60
$n = 70$	HD-SIM	0.770 (0.052)	4.15	1.33
$\sigma = 0.4$	HD-SIM-pst	0.662 (0.068)	1.84	1.64

We evaluate the methods on the accuracy of index coefficient estimation and variable selection. One performance measure is the estimation error, for which we use the sum of the absolute differences between the estimated and the true coefficients. To make sure the estimation error is independent of scaling, we formally define it for each vector $\boldsymbol{\alpha}$ as

$$\min\left(\left\|\frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|}-\boldsymbol{\alpha}^*\right\|_1, \left\|\frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|}+\boldsymbol{\alpha}^*\right\|_1\right).$$

In the rare case $\boldsymbol{\alpha} = 0$ we simply set the error to $\|\boldsymbol{\alpha}^*\|_1$. The average estimation error is reported in Table 1 for each method and each simulation setting. The corresponding standard errors are provided in parentheses. Also reported are the average number of false positives (number of noise predictors identified as signal) and the average number of false negatives (number of signal predictors identified as noise).

Results corresponding to simulation scenario 1 are provided in the top half of Table 1. The SMAVE approach does not have a $p > n$ implementation, and is excluded from the results in the corresponding settings. Putting HD-SIM-pst aside, we see that HD-SIM has the lowest estimation error in three of the four simulation settings, while SMAVE has the lowest error in the remaining setting. HD-SIM also has the smallest number of false negatives in all of the settings. On the other hand, HD-SIM selects the largest models, as seen from the false positive numbers. This illustrates the over-shrinkage phenomenon, which was discussed in the beginning of Section 5. The bottom half of Table 1 contains the results for the simulation scenario 2.

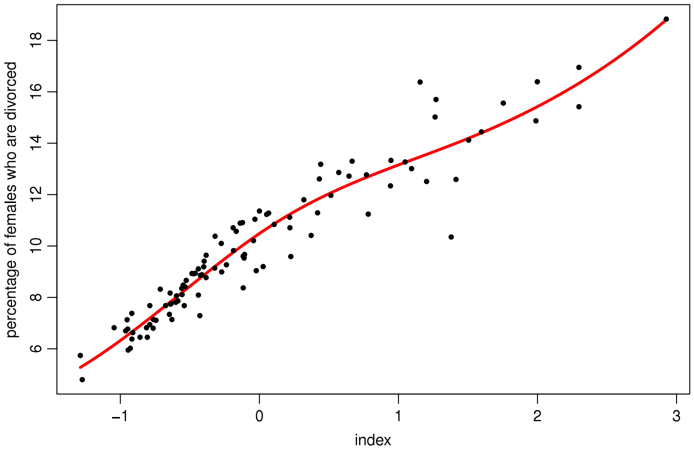

Fig. 4. Estimated regression function and observed values of the response variable are plotted against the estimated index for the Pennsylvania Communities and Crime Data.

Neither of the HD-SIM competitors performs well in this simulation setting, as evidenced by the high false negative numbers and the high estimation errors. The HD-SIM approaches perform significantly better.

The case of correlated predictors was also investigated. However, while the performance of all methods deteriorated as the correlations among the predictors increased, the results were very similar, qualitatively, to the ones presented above.

5.2. *Real data analysis*

We will consider the Communities and Crime Unnormalized Data from the UCI Machine Learning Repository, which combines socio-economic data from the 1990 Census, law enforcement data from the 1990 Law Enforcement Management and Administration Stats survey, and crime data from the 1995 FBI UCR. We will focus on explaining the response variable, *percentage of women who are divorced*, using various community characteristics, such as *percentage of population that is African American*, *percent of people in owner occupied households*, and *percent of people foreign born*. We will also use law enforcement and crime information, such as *percent of officers assigned to drug units*. In order to further explore the high-dimensional scenarios, we will look at the state-level data and examine the three largest states for which the number of predictors is at least as large as the number of observations. These states are Pennsylvania, Michigan, and Ohio.

Let us first focus on the data for the state of Pennsylvania. After removing the variables with *NA*'s and two variables directly related to the response (total and male divorce percentages), the data has 101 observation and 114 predictors. To evaluate the performance of the methods, we can randomly split the data into a training set with 70 observations, and a test set with 31 observation. Such splitting was performed one hundred times, each time fitting the methods on the training set and computing the Root Mean Square Error (RMSE) on the test set. SR-SIR approach does not produce a regression function, hence, after the index coefficients were produced on the training set, the regression function was estimated on the training set with cubic B-splines. The same number of basis functions was used for all the methods being compared. Over the 100 random splits of the data, the average RMSE for HD-SIM was 1.08, while the corresponding numbers for the three SR-SIR approaches were 1.29, 1.33 and 1.35. When compared to the best SR-SIR approach, SR-SIR AIC, the proposed HD-SIM method produced a lower RMSE in 90 out of the 100 splits of the data. HD-SIM-pst also resulted in a low RMSE of 1.07. When a paired t-test was used to compare the RMSE of the HD-SIM approaches with that of the SR-SIR approaches, each of the six p-values was below 10^{-14}.

A similar comparison of the predictive performance was also conducted for the states of Michigan and Ohio. In the case of Michigan, HD-SIM methods significantly outperformed the SR-SIR approaches, in a similar fashion to the case of Pennsylvania described above. The comparisons for Ohio were not as clear: the average RMSE for SR-SIR AIC was the lowest, followed by the HD-SIM methods, and then by SR-SIR BIC and RIC. In terms of the p-value for the paired t-test, neither of the differences between the SR-SIR and the HD-SIM approaches was statistically significant at the 5% level for the state of Ohio.

Now we will go back and look at the full Pennsylvania data set. Application of HD-SIM results in a single index model with nine predictors. The estimated regression function and the data are plotted versus the index in Fig. 4. The regression function is monotone increasing, hence the signs of the coefficients can be easily interpreted. Three predictors with the leading coefficients, *percent of people living in the same house for the last* 5 *years*, *percentage of kids in family housing with two parents* and *percentage of population that is* 16–24 *in age*, have a negative estimated relationship with the response, which is *percentage of women who are divorced*. Among the remaining predictors, *percentage of households with social security income*, *percentage of people* 16 *and over who are employed in professional services*, *percent of vacant housing that is boarded up* and *percent of vacant housing that has been vacant more than* 6 *months* have a negative relationship with the female divorce rate,

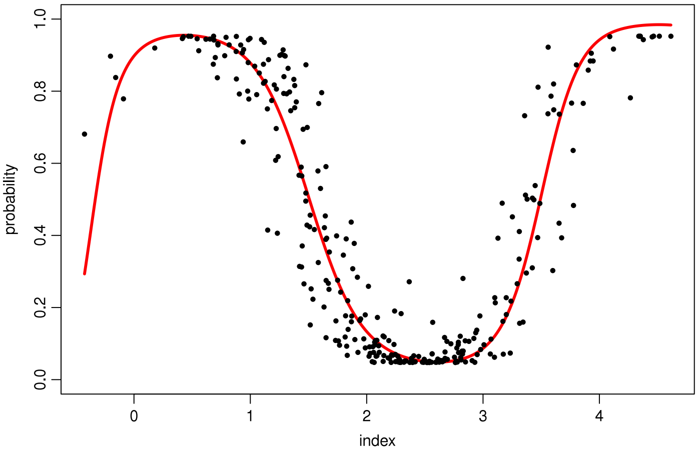

Fig. 5. Estimated success probability function and the true probabilities for the observed responses plotted against the estimated index for a Bernoulli response simulation setting.

while *per capita income for African Americans* and *per capita income for people with 'other' heritage* have a positive relationship with the response.

6. Extension to the Generalized Single Index Model setting

HD-SIM methodology can be naturally extended to the generalized response setting. For a response variable Y_i with mean μ_i we will model the relationship with the predictors as $g(\mu_i) = f^*(X_i^T \boldsymbol{\alpha}^*)$. Here g is the link function assumed to be known, with common examples including the identity link used for normal response data and the logistic link used for binary response data. For notational simplicity we will assume that g is chosen as the canonical link, though all the ideas generalize naturally to other link functions. The HD-SIM optimization problem (5) is generalized by replacing the minimization of the sum of squares with maximization of the log-likelihood function. The standard approach to finding the maximum likelihood solution in the generalized setting is iterative reweighted least squares. In particular, given the current estimate $\hat{\boldsymbol{\eta}} = \widehat{\mathbf{f}_{\widehat{\alpha}}}$, an adjusted dependent variable, $Z_i = \hat{\eta}_i + (Y_i - \hat{\mu}_i)/\hat{V}_i$, is computed, where $\hat{V}_i$ is the current estimate for the variance of Y_i. A new estimate for $\boldsymbol{\eta}$ is then produced using weighted least squares, i.e. by minimizing $\sum_i (Z_i - \eta_i)^2 \hat{V}_i$ over the allowed values of $\boldsymbol{\eta}$. This procedure is iterated until $\hat{\boldsymbol{\eta}}$ converges. In our case, the weighted least squares minimization is done over f and $\boldsymbol{\alpha}$, subject to the L_1 constraint and the identifiability condition on the index coefficients. As before, we perform the optimization by iterating the f and the $\boldsymbol{\alpha}$ updates. The f update can be done as in Section 3.1 except the basis coefficients should now be estimated using weighted least squares. After each f update we also update the values of $\hat{\eta}_i$, $\hat{\mu}_i$ and $\hat{V}_i$. The $\boldsymbol{\alpha}$ update is done by minimizing the quadratic approximation to the weighted least squares problem, which is achieved, as before, by replacing f with a first order approximation. The formulas for the $\boldsymbol{\alpha}$ update are essentially the same as in Algorithm 1, except for one small but simple modification: the two inner products in Eqs. (6) and (7) are now weighted by the $\hat{V}_i$'s.

In the following small simulation study HD-SIM approaches are compared to the SR-SIR methods. The SMAVE approach does not have a generalized response implementation and is not included. The Bernoulli responses are generated according to the following logistic regression model, where π_i is the success probability:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sin\left(X_i^T \boldsymbol{\alpha}^* \pi/2\right), \quad X_i \sim N_p(2\sqrt{17}/5, I), \; i = 1, \dots, n.$$

The true index coefficient vector, $\boldsymbol{\alpha}^*$, is again taken as $(8, 4, 2, 1, 0, \dots, 0)/\sqrt{85}$, and the number of predictors, p, is set to 100. The number of observations, n, is increased to 300, because the observed binary responses contain significantly less information than in the continuous case. Fig. 5 displays the estimated success probability curve, together with true probabilities for the observed responses, plotted against the estimated index, for one simulation run. In the one hundred simulation runs that were conducted, HD-SIM approaches outperform the SR-SIR methods. The average estimation error for HD-SIM is 0.432 (with a standard error of 0.011), while the estimation error for best SR-SIR method is 2.197 (0.196). The average number of false positives for HD-SIM is 5.05, while the average number of false negatives is 0.43. The corresponding numbers for the best SR-SIR approach are 8.99 and 2.46. The estimation errors for the other SR-SIR methods are very high, while the value for HD-SIM-pst is even lower than that of HD-SIM.

7. Conclusions

The paper presents a novel approach, HD-SIM, for efficiently estimating all the components of the single index model and performing variable selection simultaneously. The method is designed for the high-dimensional setting, which includes situations where p is much larger than n. The level of regularization is controlled by a tuning parameter, and an algorithm is presented for computing a solution path with respect to this parameter in a computationally efficient manner. Asymptotic theory is developed for the proposed estimates of the regression function and the index coefficients in the high-dimensional setting. A simulation study and analysis of real data demonstrate that the new estimator performs very well versus various natural competitors.

One potential extension of the HD-SIM methodology would be to allow the regression function to have multiple components, each being a function of its own index. The new method would need to encourage sparsity in both the index coefficients and the functional components. It would be interesting to see whether ideas from some of the recent approaches for fitting high-dimensional additive models could be integrated with the HD-SIM methodology to achieve the aforementioned goal.

Acknowledgment

This work was partially supported by National Science Foundation Grant DMS-1209057.

Appendix A. Theoretical assumptions

Given a real valued function f, a vector $\boldsymbol{\alpha} \in \mathbb{R}^p$ and an index j, we will define functions $f_{\boldsymbol{\alpha}} : x \mapsto f(x^T\boldsymbol{\alpha})$ and $f_j : x \mapsto f(x_j)$ for $x \in \mathbb{R}^p$. Let P_n denote the empirical probability measure associated with the predictor vectors, $X_1, \ldots, X_n$. For simplicity of the notation we will write $\|\cdot\|_n$ instead of $\|\cdot\|_{L_2(P_n)}$. Given a B-spline basis, $\{b_1, \ldots, b_{d_n}\}$, presented in Section 3.1 for the class $\mathcal{F}_n(\boldsymbol{\alpha})$, let $B_{\boldsymbol{\alpha}}$ denote the vector valued function $x \mapsto \big(b_1(x^T\boldsymbol{\alpha}), \ldots, b_{d_n}(x^T\boldsymbol{\alpha})\big)$. Define $\boldsymbol{\beta}^* = \arg\min_{\beta} \|B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta} - f^*_{\boldsymbol{\alpha}^*}\|_n$, so that $B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*$ is the B-spline representation of the true regression function. Note that $\boldsymbol{\beta}^*$ depends on n, but we will omit the subscript for the simplicity of the notation. To protect against some extreme scenarios in the construction of the B-spline basis, we will select a very small positive constant ϵ, and for all $\boldsymbol{\alpha}$ restrict the length of the interval associated with $\mathcal{F}_n(\boldsymbol{\alpha})$ from dropping below ϵ. Recall that $\boldsymbol{\alpha}^*$ is scaled to satisfy $\|\boldsymbol{\alpha}^*\|_1 = 1$, and that predictors are treated as deterministic. Write Q_n for the empirical distribution associated with the true index values $X_1^T\boldsymbol{\alpha}^*, \ldots, X_n^T\boldsymbol{\alpha}^*$. Let $\mathcal{A}^*$ denote the index sets for the nonzero coefficients of $\boldsymbol{\alpha}^*$ and define $\mathcal{S} = \{(\boldsymbol{\alpha}, \boldsymbol{\beta}) : \ \|\boldsymbol{\alpha}\|_1 \le \|\boldsymbol{\alpha}^*\|_1, \ \alpha_m = \alpha^*_m \text{ for some } m \in \mathcal{A}^*\}$. The following assumptions are used in Section 4. A *universal constant* should be interpreted as a constant that does not depend on n or any of the other parameters that appear in the corresponding expression.

A1. $|X_{ij}| \le C_1$ for all i and j, where C_1 is a universal constant.
A2. The true regression function, f^*, is twice continuously differentiable.
A3. $\sup_u \big|Q_n(-\infty, u] - Q(-\infty, u]\big| = o\big(d_n^{-1}\big)$ for some probability distribution Q with bounded support and a positive continuous density.
A4. Errors ε_i are independent and uniformly subgaussian.
A5. $\|\boldsymbol{\beta}^*\|_\infty \le M$ and $\mathcal{F}_n(\boldsymbol{\alpha}) = \{\sum_{j=1}^{d_n} \beta_j b_j, \ \|\boldsymbol{\beta}\|_\infty \le M\}$ for some universal M.
A6. If $\|B_{\boldsymbol{\alpha}_n}\boldsymbol{\beta}_n - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n \to 0$ for a sequence $(\boldsymbol{\alpha}_n, \boldsymbol{\beta}_n)$ in the set $\mathcal{S}$, then $\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*\| \to 0$ and $\|B_{\boldsymbol{\alpha}^*}(\boldsymbol{\beta}_n - \boldsymbol{\beta}^*)\|_n \to 0$.
A7. $\min_{j\in\mathcal{A}^*} \min_{\{f\in\mathcal{F}_n(j)\}} \|B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^* - f_j\|_n^2 + \epsilon \le \min_{k\notin\mathcal{A}^*} \min_{\{f\in\mathcal{F}_n(k)\}} \|B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^* - f_k\|_n^2$, for some positive ϵ and all sufficiently large n.

Assumptions A1–A3 are standard in the spline estimation literature, while A4 is typical for deriving asymptotics of nonparametric least-squares estimators. Assumption A5 is technical and could be relaxed at the cost of extra conditions on the global behavior of the sum of squares function. It is stated in the present form for simplicity of the exposition. Unconstrained form of A6 is required for establishing consistency in the classical setting, where the number of predictors is fixed. Note that in the special case of linear regression, where $B_{\boldsymbol{\alpha}}\boldsymbol{\beta}$ is replaced by $X\boldsymbol{\alpha}$, assumption A6 follows from the *restricted eigenvalue assumption* of [2], which is required for the Lasso estimation theory. Assumption A7 means that the best signal predictor does a better job marginally approximating the true regression function, $f^*_{\boldsymbol{\alpha}^*}$, than the noise predictors. It implies that, with probability tending to one, $\min_{j\in\mathcal{A}^*} \min_{\{f\in\mathcal{F}_n(j)\}} \|y - f_j\|_{\underline{n}}^2 < \min_{k\notin\mathcal{A}^*} \min_{\{f\in\mathcal{F}_n(k)\}} \|y - f_k\|_n^2$, which means $\widehat{m} \in \mathcal{A}^*$. It follows from the proof of Theorem 1 that consistency of $\widehat{f_{\widehat{\alpha}}}$ holds even without A6 and A7, but these assumptions are required for the consistency of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{f}$.

To simplify the presentation of the remaining assumption, B, we will suppose that there exists some arbitrarily small L_∞ neighborhood of the true index vector, $X\boldsymbol{\alpha}^*$, such that the same sequence of knots is used in the construction of the B-spline basis for all the index vectors in this neighborhood. Set $\Delta\boldsymbol{\alpha} = \boldsymbol{\alpha} - \boldsymbol{\alpha}^*$ and $\Delta\boldsymbol{\gamma} = d_n^{-1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$, and define the cone $\mathcal{C} = \{(\Delta\boldsymbol{\alpha}, \Delta\boldsymbol{\gamma}) : \|\Delta\boldsymbol{\alpha}_{\mathcal{A}^{*c}}\|_1 \le \|\Delta\boldsymbol{\alpha}_{\mathcal{A}^*}\|_1\}$. Here we use $\Delta\boldsymbol{\alpha}_{\mathcal{A}}$ to denote the sub-vector of $\Delta\boldsymbol{\alpha}$ identified by $\mathcal{A}$. To derive the rate of convergence we will need to impose a regularity condition controlling the behavior of the sum of squares near the minimum. Unconstrained assumptions of this form are also imposed in the classical setting. We will view $\|B_{\boldsymbol{\alpha}}\boldsymbol{\beta} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n^2$ as a function of $(\Delta\boldsymbol{\alpha}, \Delta\boldsymbol{\gamma})$ and let Σ_n denote the corresponding second derivative matrix at zero.

B. The restricted eigenvalues of Σ_n, corresponding to the cone $\mathcal{C}$, are bounded away from zero uniformly in n.

In the linear regression case, assumption B becomes the restricted eigenvalue assumption of [2], mentioned earlier. Also note that for $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ assumption B follows from A3. Indeed, by the properties of B-spline functions [40] there exists a positive constant c, such that $\|B_{\boldsymbol{\alpha}^*}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_n^2 > c\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 d_n^{-1}$.

Appendix B. Proof of Theorem 1

The argument will be conducted on the event $\alpha^*_{\widehat{m}} \neq 0$, which holds with probability tending to one, as pointed out in Appendix A. We will set the HD-SIM tuning parameter, t, equal to $\|\boldsymbol{\alpha}^*\|_1/|\alpha^*_{\widehat{m}}|$. Recall that we focus on the scaling of $\widehat{\boldsymbol{\alpha}}$ that achieves $\widehat{\alpha}_{\widehat{m}} = \alpha^*_{\widehat{m}}$. Note that the rescaled $\widehat{\boldsymbol{\alpha}}$ also satisfies $\|\widehat{\boldsymbol{\alpha}}\|_1 = \|\boldsymbol{\alpha}^*\|_1 = 1$.

Our proof of Theorems 1 and 2 will use the empirical process approach to the asymptotics of nonlinear least-squares estimation. This approach is discussed in, for example, [30,25]. Consider the following identity:

$$\|y - \widehat{f}_{\widehat{\boldsymbol{\alpha}}}\|_n^2 - \|y - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n^2 = \|\widehat{f}_{\widehat{\boldsymbol{\alpha}}} - f^*_{\boldsymbol{\alpha}^*}\|_n^2 - \|B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^* - f^*_{\boldsymbol{\alpha}^*}\|_n^2 - 2(\varepsilon, \widehat{f}_{\widehat{\boldsymbol{\alpha}}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*)_n. \tag{9}$$

Note that the left-hand side of (9) is non-positive by the definition of $\widehat{\boldsymbol{\alpha}}$. Consequently,

$$\|\widehat{f}_{\widehat{\boldsymbol{\alpha}}} - f^*_{\boldsymbol{\alpha}^*}\|_n^2 \leq \|B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^* - f^*_{\boldsymbol{\alpha}^*}\|_n^2 + 2(\varepsilon, \widehat{f}_{\widehat{\boldsymbol{\alpha}}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*)_n. \tag{10}$$

Observe that

$$(\varepsilon, \widehat{f}_{\widehat{\boldsymbol{\alpha}}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*)_n \leq \sup_{g \in \mathcal{G}_n}(\varepsilon, g)_n, \tag{11}$$

where $\mathcal{G}_n = \{B_{\boldsymbol{\alpha}}\boldsymbol{\beta} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*, \|\boldsymbol{\alpha}\|_1 \leq \|\boldsymbol{\alpha}^*\|_1, \|\boldsymbol{\beta}\|_\infty \leq M\}$. Given a pseudo-metric space $(\mathcal{X}, d)$, we will use $N(u, \mathcal{X}, d)$ to denote the smallest number N, such that N balls of d-radius u can cover $\mathcal{X}$. We will also write $H(u, \mathcal{X}, d)$ for $\log N(u, \mathcal{X}, d)$. We can bound the right-hand side in (11) by controlling the entropy integral $\int H^{1/2}(u, \mathcal{G}_n, \|\cdot\|_n)du$. We will use the following result, which is proved in Appendix D.

Lemma 1. *There exists a positive universal constant C for which*

(i) $\sup_{g \in \mathcal{G}_n}\|g\|_n \leq Cd_n$

(ii) $\int_{n^{-1/2}}^{Cd_n} H^{1/2}(u, \mathcal{G}_n, \|\cdot\|_n)du \lesssim d_n \log n\sqrt{\log(np_n)}$.

Lemma 1 and a maximal inequality for weighted sums of subgaussian variables, e.g. Corollary 8.3 of [30], give $\sup_{g \in \mathcal{G}_n}(\varepsilon, g)_n = O_p(n^{-1/2}d_n \log n\sqrt{\log(np_n)})$. By inequality (11) the last expression provides a stochastic bound for the second term on the right-hand side of (10). The first term on the right-hand side of (10) is the cubic B-spline approximation error, which is $O(d_n^{-2})$ under assumptions A2 and A3 in Appendix A. Combining the two terms yields the bound $\|\widehat{f}_{\widehat{\boldsymbol{\alpha}}} - f^*_{\boldsymbol{\alpha}^*}\|_n = O_p(n^{-1/2}d_n \log n\sqrt{\log(np_n)} + d_n^{-2})$, which simplifies to $o_p(1)$ by the assumption on the rates of growth for p_n and d_n. Applying the approximation error bound again, we derive

$$\|B_{\widehat{\boldsymbol{\alpha}}}\widehat{\boldsymbol{\beta}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n = o_p(1). \tag{12}$$

Consistency of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{f}$ is then a direct consequence of assumption A6.

Appendix C. Proof of Theorems 2 and 3

As in the proof of Theorem 1, we will conduct the argument on the event $\alpha^*_{\widehat{m}} \neq 0$ and focus on the choice of t and the scaling of $\widehat{\boldsymbol{\alpha}}$ that guarantee $\widehat{\alpha}_{\widehat{m}} = \alpha^*_{\widehat{m}}$ and $\|\widehat{\boldsymbol{\alpha}}\|_1 = \|\boldsymbol{\alpha}^*\|_1 = 1$.

Combine inequality $\|\widehat{f}_{\widehat{\boldsymbol{\alpha}}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n^2 \leq 2\|\widehat{f}_{\widehat{\boldsymbol{\alpha}}} - f^*_{\boldsymbol{\alpha}^*}\|_n^2 + 2\|B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^* - f^*_{\boldsymbol{\alpha}^*}\|_n^2$ with inequality (10) to derive

$$\|\widehat{f}_{\widehat{\boldsymbol{\alpha}}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n^2 \leq 4\|B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^* - f^*_{\boldsymbol{\alpha}^*}\|_n^2 + 4(\varepsilon, \widehat{f}_{\widehat{\boldsymbol{\alpha}}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*)_n. \tag{13}$$

On the event $A_n = \{(\varepsilon, \widehat{f}_{\widehat{\boldsymbol{\alpha}}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*)_n < \|B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^* - f^*_{\boldsymbol{\alpha}^*}\|_n^2\}$ the above inequality yields

$$\|\widehat{f}_{\widehat{\boldsymbol{\alpha}}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n < 2\sqrt{2}\|B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^* - f^*_{\boldsymbol{\alpha}^*}\|_n = O(d_n^{-2}), \tag{14}$$

where the stochastic bound comes from the cubic B-spline approximation error.

The following argument will be conducted on the event A_n^c. By inequality (13) we have

$$\|\widehat{f}_{\widehat{\boldsymbol{\alpha}}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n^2 \leq 8(\varepsilon, \widehat{f}_{\widehat{\boldsymbol{\alpha}}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*)_n. \tag{15}$$

We will apply the peeling device and proceed in a similar fashion to the proof of Theorem 9.1 in [30]. We will use the following result, which is proved in Appendix D. Define $\mathcal{G}_n(\delta) = \{g \in \mathcal{G}_n, \|g\|_n \leq \delta, \alpha_m = \alpha^*_m \text{ for some } m \in \mathcal{A}^*\}$.

Lemma 2. *There exists a positive constant* τ, *such that for all* $\delta \le \tau$,

$$\int_{n^{-1/2}}^{\delta} \sqrt{\log N(u, \mathcal{G}_n(\delta), \|\cdot\|_n)}du \lesssim s_n^{1/2}(\delta + \delta^2 d_n^{3/2}) \log n \sqrt{\log p_n} + \delta d_n^{1/2}\sqrt{\log n}. \tag{16}$$

Let S_n be the largest integer such that $2^{S_n+1}\delta_n \le \tau$; sequence δ_n will be defined later. Let $q_n = P(\|\widehat{f_{\alpha}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n \ge \tau/2)$ and note that $q_n = o(1)$, according to display (12). Taking advantage of inequality (15), we derive the following bounds:

$$\begin{aligned} P(\|\widehat{f_{\alpha}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n > \delta_n) &\le q_n + \sum_{s=0}^{S_n} P(2^s\delta_n < \|\widehat{f_{\alpha}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n \le 2^{s+1}\delta_n) \\ &\le q_n + \sum_{s=0}^{S_n} P(\sup_{g\in\mathcal{G}_n(2^{s+1}\delta_n)} (\varepsilon, g)_n > 2^{2s-3}\delta_n^2) = q_n + \sum_{s=0}^{S_n} P_s, \end{aligned}$$

we can again apply Corollary 8.3 of [30] to bound each P_s. Let c be a universal constant and take $\delta_n = cn^{-1/2}(s_n^{1/2}\log n\sqrt{\log p_n} + d_n^{1/2}\sqrt{\log n})$. Constant c needs to be sufficiently large to ensure that $\sqrt{n}\delta^2$ is larger than a fixed multiple of the right hand side in (16) for all $\delta \ge \delta_n$. Such a constant exists, because $s_n^{1/2}d_n^{3/2}\log n\sqrt{\log p_n} = o(\sqrt{n})$ is one of the assumptions of the theorem. It follows that for some constants c_1 and c_2,

$$\sum_{s=0}^{S_n} P_s \le \sum_{s=0}^{S_n} c_1 \exp(-n2^{4s-6}\delta_n^4/c_1^2 2^{2s+2}\delta_n^2) \le c_2\exp(-n\delta_n^2/c_2) \to 0,$$

as n goes to infinity. Hence, $\|\widehat{f_{\alpha}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n = O\big(n^{-1/2}(s_n^{1/2}\log n\sqrt{\log p_n} + d_n^{1/2}\sqrt{\log n})\big)$, with probability tending to one. Recall that we restricted our attention to the events A_n^c. Taking into account events A_n, we have, in view of (14),

$$\|\widehat{f_{\alpha}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n = O\left(n^{-1/2}\big(s_n^{1/2}\log n\sqrt{\log p_n} + d_n^{1/2}\sqrt{\log n}\big) + d_n^{-2}\right) = O(r_n), \tag{17}$$

with probability tending to one. Assumption B implies that the above bound also holds for $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|$ and $\|B_{\boldsymbol{\alpha}^*}\widehat{\boldsymbol{\beta}} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n$. This completes the proof of Theorem 2.

The bound on $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|$ implies that, with probability tending to one,

$$\sum_{j\notin\mathcal{A}^*} |\widetilde{\alpha}_j|^2 \le \sum_{j\notin\mathcal{A}^*} |\widehat{\alpha}_j - \alpha_j^*|^2 = O(s_n q_n^2). \tag{18}$$

Use bound (18) and inequalities $|\widetilde{\alpha}_j|^2 > q_n^2$, for $j \in \widetilde{\mathcal{A}}_n$, to deduce $|(\mathcal{A}^*)^c \cap \widetilde{\mathcal{A}}_n| = O(s_n)$. This implies $|\widetilde{\mathcal{A}}_n| \le |\mathcal{A}^*| + |(\mathcal{A}^*)^c \cap \widetilde{\mathcal{A}}_n| = O(s_n)$. Also note that

$$\sum_{j\in\mathcal{A}^*} |\widetilde{\alpha}_j - \alpha_j^*|^2 \le \sum_{j\in\mathcal{A}^*}(q_n^2 + |\widehat{\alpha}_j - \alpha_j^*|^2) = O(s_n q_n^2).$$

The above bound, together with (18), yields the error bound in Theorem 3.

Appendix D. Proof of Lemmas 1 and 2

Lemma 1. For $\boldsymbol{\alpha} \in \mathbb{R}^p$ we will use $\boldsymbol{\alpha}^T(\cdot)$ to denote the function $x \mapsto \boldsymbol{\alpha}^T x$, defined on $\mathbb{R}^p$. Given vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ and the B-spline basis, $\{b_1, \ldots, b_{d_n}\}$, for the class $\mathcal{F}_n(\boldsymbol{\alpha}_1)$, we will define $\mathbf{b}_{\boldsymbol{\alpha}_1}$ as a row vector valued function $\big(b_1, \ldots, b_{d_n}\big)$. We will write $a(\boldsymbol{\alpha}) = \min_i X_i^T\boldsymbol{\alpha}$, $b(\boldsymbol{\alpha}) = \max_i X_i^T\boldsymbol{\alpha}$, and $e(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = |a(\boldsymbol{\alpha}_1) - a(\boldsymbol{\alpha}_2)| + |b(\boldsymbol{\alpha}_1) - b(\boldsymbol{\alpha}_2)|$. Recall that $b(\boldsymbol{\alpha}) - a(\boldsymbol{\alpha})$ is not allowed to drop below some small positive ϵ. By the properties of B-splines,

$$\begin{aligned} &\|B_{\boldsymbol{\alpha}}\boldsymbol{\beta}\|_\infty \le \|\boldsymbol{\beta}\|_\infty \\ &\|B_{\boldsymbol{\alpha}_2}\boldsymbol{\beta} - B_{\boldsymbol{\alpha}_1}\boldsymbol{\beta}\|_n \quad = \|\mathbf{b}_{\boldsymbol{\alpha}_1}(c_1 + c_2\boldsymbol{\alpha}_2^T\cdot)\boldsymbol{\beta} - \mathbf{b}_{\boldsymbol{\alpha}_1}(\boldsymbol{\alpha}_1^T\cdot)\boldsymbol{\beta}\|_n, \end{aligned} \tag{19}$$

where $\max(|c_1|, ||c_2 - 1|) \lesssim e(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$. Let $h = \mathbf{b}_{\boldsymbol{\alpha}_1}(\boldsymbol{\alpha}_1^T\cdot)\boldsymbol{\beta}$ and note that $|h(z_2) - h(z_1)| \lesssim \|\boldsymbol{\beta}\|_\infty d_n|z_2 - z_1|$ by the properties of the B-spline derivatives. Consequently,

$$\|B_{\boldsymbol{\alpha}_2}\boldsymbol{\beta} - B_{\boldsymbol{\alpha}_1}\boldsymbol{\beta}\|_n \lesssim \|\boldsymbol{\beta}\|_\infty d_n\Big(\|(\boldsymbol{\alpha}_2 - \boldsymbol{\alpha}_1)^T(\cdot)\|_n + e(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)\Big). \tag{20}$$

Using the bound $\max_{i\le n}|X_i\boldsymbol{\alpha}^T| \lesssim 1$, together with displays (19) and (20), we derive

$$\begin{aligned} \|B_{\boldsymbol{\alpha}}\boldsymbol{\beta} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n &\le \|B_{\boldsymbol{\alpha}^*}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_n + \|B_{\boldsymbol{\alpha}}\boldsymbol{\beta} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}\|_n \\ &\lesssim \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\infty + \|\boldsymbol{\beta}\|_\infty d_n\Big(\|(\boldsymbol{\alpha}_2 - \boldsymbol{\alpha}_1)^T(\cdot)\|_n + e(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)\Big) \lesssim d_n, \end{aligned}$$

for $(B_{\boldsymbol{\alpha}}\boldsymbol{\beta} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*) \in \mathcal{G}_n$. This establishes part (i) of the lemma.

Consider $g_k = B_{\boldsymbol{\alpha}_k}\boldsymbol{\beta}_k - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*$, such that $g_k \in \mathcal{G}_n$ for $k \in \{1, 2\}$. We can write $\|g_2 - g_1\|_n \le \|B_{\boldsymbol{\alpha}_2}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)\|_n + \|B_{\boldsymbol{\alpha}_2}\boldsymbol{\beta}_1 - B_{\boldsymbol{\alpha}_1}\boldsymbol{\beta}_1\|_n$, which then, by inequalities (19) and (20), gives us

$$\|g_2 - g_1\|_n \lesssim \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|_\infty + Md_n\Big(\|(\boldsymbol{\alpha}_2 - \boldsymbol{\alpha}_1)^T(\cdot)\|_n + e(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)\Big).$$

Observe that $H(u, \{\boldsymbol{\beta} \in \mathbb{R}^{d_n} : \|\boldsymbol{\beta}\|_\infty \le M\}, \|\cdot\|_\infty) \lesssim d_n[1 + \log(1/u)]$. Lemma 2.6.11 in [31] yields, for $u \le Cd_n$,

$$H(u, \{x \mapsto Md_n\boldsymbol{\alpha}^T x, \boldsymbol{\alpha} \in \mathbb{R}^p, \|\boldsymbol{\alpha}\|_1 \le 1\}, \|\cdot\|_n) \lesssim d_n^2 u^{-2}[1 + \log(d_n p_n)].$$

Note that $N(u, \{Md_n a(\boldsymbol{\alpha}) : \|\boldsymbol{\alpha}\|_1 \le 1\}, |\cdot|) \lesssim d_n u^{-1}$, and the same bound holds for the class that involves $b(\boldsymbol{\alpha})$. Consequently,

$$H^{1/2}(u, \mathcal{G}_n, \|\cdot\|_n) \lesssim d_n^{1/2}\sqrt{1 + \log(1/u)} + d_n u^{-1}\sqrt{1 + \log(np_n)},$$

from which statement (ii) of the lemma follows after the integration.

Lemma 2. For every $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ we will write $\Delta\boldsymbol{\alpha}$ for $\boldsymbol{\alpha} - \boldsymbol{\alpha}^*$ and $\Delta\boldsymbol{\beta}$ for $\boldsymbol{\beta} - \boldsymbol{\beta}^*$. Note that $\|\boldsymbol{\alpha}\|_1 \le \|\boldsymbol{\alpha}^*\|_1$ implies $\|\Delta\boldsymbol{\alpha}_{\mathcal{A}^{*c}}\|_1 \le \|\Delta\boldsymbol{\alpha}_{\mathcal{A}^*}\|_1$. We will take τ small enough to ensure that on $\mathcal{G}_n(\tau)$ function $\|B_{\boldsymbol{\alpha}}\boldsymbol{\beta} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*\|_n^2$ can be bounded below by a positive multiple of $\|\Delta\boldsymbol{\alpha}\|^2 + \|\Delta\boldsymbol{\beta}\|^2 d_n^{-1}$. Assumption B guarantees that such a τ can be found. Take $\delta \le \tau$ and consider a function $(B_{\boldsymbol{\alpha}}\boldsymbol{\beta} - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*) \in \mathcal{G}_n(\delta)$. There exists a positive universal constant κ for which $\|\Delta\boldsymbol{\beta}\|_\infty \le \delta\kappa d_n^{1/2}$ and $\|\Delta\boldsymbol{\alpha}\| \le \kappa\delta$. Properties of B-spline derivatives yield

$$\|(B_{\boldsymbol{\alpha}}\boldsymbol{\beta})'\|_\infty \lesssim \|(B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta})'\|_\infty \lesssim 1 + \|\Delta\boldsymbol{\beta}\|_\infty d_n \lesssim 1 + \delta d_n^{3/2}. \tag{21}$$

Suppose $(B_{\boldsymbol{\alpha}_k}\boldsymbol{\beta}_k - B_{\boldsymbol{\alpha}^*}\boldsymbol{\beta}^*) \in \mathcal{G}_n(\delta)$ for $k \in \{1, 2\}$. Inequalities (19)–(21) give

$$\begin{aligned}\|B_{\boldsymbol{\alpha}_2}\boldsymbol{\beta}_2 - B_{\boldsymbol{\alpha}_1}\boldsymbol{\beta}_1\|_n &\le \|B_{\boldsymbol{\alpha}_2}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)\|_n + \|B_{\boldsymbol{\alpha}_2}\boldsymbol{\beta}_1 - B_{\boldsymbol{\alpha}_1}\boldsymbol{\beta}_1\|_n \\ &\lesssim \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|_\infty + \|(B_{\boldsymbol{\alpha}_1}\boldsymbol{\beta}_1)'\|_\infty\Big(\|(\boldsymbol{\alpha}_2 - \boldsymbol{\alpha}_1)^T(\cdot)\|_n + e(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)\Big) \\ &\lesssim \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|_\infty + (1 + \delta d_n^{3/2})\Big(\|(\boldsymbol{\alpha}_2 - \boldsymbol{\alpha}_1)^T(\cdot)\|_n + e(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)\Big).\end{aligned}$$

Define $S_n(\delta) = \{\Delta\boldsymbol{\beta},\ \|\Delta\boldsymbol{\beta}\|_\infty \le \delta(\kappa d_n)^{1/2}\}$ and note that $H(u, S_n(\delta), \|\cdot\|_\infty) \lesssim d_n \log(\delta d_n^{1/2}/u)$. Thus,

$$\int_{n^{-1/2}}^{\delta} H^{1/2}(u, S_n(\delta), \|\cdot\|_\infty)du \lesssim \delta d_n^{1/2}\sqrt{\log n}. \tag{22}$$

Define $\mathcal{A}_n(\delta) = \{x \mapsto (1 + \delta d_n^{3/2})\Delta\boldsymbol{\alpha}^T x,\ \|\Delta\boldsymbol{\alpha}\| \le \kappa^{-1/2}\delta, \|\boldsymbol{\alpha}\|_1 \le \|\boldsymbol{\alpha}^*\|_1\}$. Observe that $\|\boldsymbol{\alpha}\|_1 \le \|\boldsymbol{\alpha}^*\|_1$ implies $\|\Delta\boldsymbol{\alpha}\|_1 \le 2\|\Delta\boldsymbol{\alpha}_{\mathcal{A}^*}\|_1 \le 2s_n^{1/2}\|\Delta\boldsymbol{\alpha}\|$. It follows that $A_n(\delta)$ is a subset of $\tilde{\mathcal{A}}_n(\delta) = \{x \mapsto (1 + \delta d_n^{3/2})\Delta\boldsymbol{\alpha}^T x,\ \|\Delta\boldsymbol{\alpha}\|_1 \le C_3 s_n^{1/2}\delta\}$. Another application of Lemma 2.6.11 in [31] yields, for $u \le \tau$,

$$H\Big(u, \tilde{\mathcal{A}}_n(\delta), \|\cdot\|_n\Big) \lesssim (1 + \delta d_n^{3/2})^2(s_n^{1/2}\delta/u)^2[1 + \log p_n].$$

Consequently,

$$\int_{n^{-1/2}}^{\delta} H^{1/2}\Big(u, \tilde{\mathcal{A}}_n(\delta), \|\cdot\|_n\Big) \lesssim s_n^{1/2}\delta(1 + \delta d_n^{3/2}) \log n\sqrt{\log p_n}. \tag{23}$$

Define $\mathcal{C}_n(\delta) = \{(1 + \delta d_n^{3/2})a(\boldsymbol{\alpha}) : \|\Delta\boldsymbol{\alpha}\| \le \kappa^{-1/2}\delta, \|\boldsymbol{\alpha}\|_1 \le \|\boldsymbol{\alpha}^*\|_1\}$. Note that $N(u, \mathcal{C}_n(\delta), |\cdot|) \lesssim s_n^{1/2}\delta(1 + \delta d_n^{3/2})u^{-1}$. Hence, for $\delta \le \tau$,

$$\int_{n^{-1/2}}^{\delta} H^{1/2}\Big(u, \mathcal{C}_n(\delta), |\cdot|\Big) \lesssim \delta(\sqrt{\log n} + \sqrt{\log s_n}). \tag{24}$$

The same bound holds for the class that involves $b(\boldsymbol{\alpha})$.

The result of the lemma follows from Eqs. (22)–(24).

Appendix E. Proof of Propositions 1 and 2

Proposition 1. For concreteness, let $\alpha^* > 0$. The proof for the case $\alpha^* < 0$ is essentially the same. Function F_λ is strictly decreasing on $(-\infty, 0)$ and increasing on (α^*, ∞). Hence, all of its local minima lie in $[0, \alpha^*]$. Note that $G'(\alpha) < 0$ for $\alpha < \alpha^*$ and $G'(\alpha) > 0$ for $\alpha > \alpha^*$. The zero subgradient condition for F_λ is satisfied at a positive α if $G'(\alpha) + \lambda\alpha = 0$, and it is satisfied at zero if $|G'(0)| \le \lambda$. Define $a = \arg\min_{\alpha \ge 0} G'(\alpha)$, and note that a is positive by the assumption $|G'(0)| < \sup_{(0,\alpha^*)}|G'|$. For $\lambda > |G'(a)|$, the subgradient condition is satisfied only at zero, hence zero is the unique global minimum of F_λ. For

$\lambda \in (|G'(0)|, |G'(a)|)$, the subgradient condition is satisfied at zero, at a point in $(0, a)$ and at a point in (a, ∞). It follows that F_λ has a local minimum at 0, a local maximum in $(0, a)$ and a local minimum in $(a, \alpha^*]$. For $\lambda < |G'(0)|$, the subgradient condition can only be satisfied at a point in $(a, \alpha^*]$. This point is the global minimum of F_λ. Hence, the statement of the proposition is valid for $\lambda_1 = |G'(0)|$ and $\lambda_3 = |G'(a)|$. Note that function $D(\lambda) = F_\lambda(0) - \min_{a\in[a,\alpha^*]} F_\lambda(a)$ is continuous. Also note that $D(\lambda_1) \geq 0$, while $D(\lambda_3) \leq 0$. Thus, there exists a λ_2 in $[\lambda_1, \lambda_3]$, for which $D(\lambda_2) = 0$, and the value of F_λ at its two local minima is the same. However, when $\lambda = \lambda_1$ or $\lambda = \lambda_3$, the subgradient condition for F_λ is satisfied at only two points, which implies that F_λ can have at most one local minimum. Thus, $\lambda_2 \in (\lambda_1, \lambda_3)$.

Proposition 2. Write $\widehat{\boldsymbol{\alpha}}(t)$ for the solution to the optimization problem (4). Define $H_t(\boldsymbol{\alpha}) = \left\|\mathbf{Y} - \mathbf{f}_{t\boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|_1}\right\|$. Note that $\widehat{\boldsymbol{\alpha}}(t)/t$ is the global minimum of $H_t(\boldsymbol{\alpha})$ over the set $A_1 = \{\boldsymbol{\alpha}, \|\boldsymbol{\alpha}\|_1 = 1\}$. Continuity of f implies equicontinuity of $H_t(\boldsymbol{\alpha})$, as a function of t, over the set A_1. Consequently, for each $t_0 \in (T_1, T_2)$, functions $H_t(\boldsymbol{\alpha})$ converge to $H_{t_0}(\boldsymbol{\alpha})$, as $t \to t_0$, uniformly over $\|\boldsymbol{\alpha}\|_1 = 1$. It follows that $\arg\min_{A_1} H_t$ converges to $\arg\min_{A_1} H_{t_0}$. Thus, $\widehat{\boldsymbol{\alpha}}(t)/t \to \widehat{\boldsymbol{\alpha}}(t_0)/t_0$ as $t \to t_0$, which implies $\widehat{\boldsymbol{\alpha}}(t) \to \widehat{\boldsymbol{\alpha}}(t_0)$. Cases $t_0 = T_1$ and $t_0 = T_2$ can be handled analogously.

Appendix F. Derivation of formula (7)

Suppose first that we start with $\alpha_j \neq 0$. Then we must have $\Delta_L = -\Delta_j \text{sign}(\alpha_L\alpha_j)$, which can be written as $\Delta_L = -\Delta_j S_{jL}$, in order to maintain the L_1 constraint. For example, if both coefficients are positive, then $\Delta_L = -\Delta_j$. After we express Δ_L in terms of Δ_j, we differentiate the objective, $\frac{1}{2}\left\|\mathbf{R} - \mathbf{f}' * (\mathbf{X}_L\Delta_L + \mathbf{X}_j\Delta_j)\right\|^2$, with respect to Δ_j, and obtain $\left(\mathbf{f}' * (S_{jL}\mathbf{X}_L - \mathbf{X}_j)\right)^T \left(\mathbf{R} - \Delta_j\mathbf{f}' * (\mathbf{X}_j - S_{jL}\mathbf{X}_L)\right) = \left(\nabla_j - S_{jL}\nabla_L\right) + \Delta_j \left\|\mathbf{f}' * \left(\mathbf{X}_j - S_{jL}\right)\right\|^2$. Setting this expression to zero and solving for Δ_j leads directly to (7). Now suppose that we start with $\alpha_j = 0$. In this case, the L_1 constraint implies $\Delta_L = \Delta_j \text{sign}(\alpha_L\nabla_j)$, which simplifies to $\Delta_L = -\Delta_j S_{jL}$. For example, if Δ_L is positive and Δ_j is about to become positive, i.e. $\nabla_j < 0$ and $|\nabla_j| > |\nabla_L|$, then $\Delta_L = -\Delta_j$. Arguing as before, we arrive at (7).

References

[1] R.E. Bellman, Adaptive Control Processes, Princeton University Press, 1961.
[2] P. Bickel, Y. Ritov, A. Tsybakov, Simultaneous analysis of lasso and dantzig selector, Ann. Statist. 37 (2009) 1705–1732.
[3] H. Bondell, L. Li, Shrinkage inverse regression estimation for model free variable selection, J. R. Stat. Soc. Ser. B 71 (2009) 287–299.
[4] E. Candes, T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n (with discussion), Ann. Statist. 35 (6) (2007) 2313–2351.
[5] N.H. Choi, W. Li, J. Zhu, Variable selection with the strong heredity constraint and its oracle property, J. Amer. Statist. Assoc. 105 (2010) 354–364.
[6] R. Cook, Regression Graphics: Ideas for Studying Regressions Throgh Graphics, Wiley, New York, 1998.
[7] X. Cui, W. Härdle, L. Zhu, The EFM approach for single-index models, Ann. Statist. 39 (2011) 1658–1688.
[8] J. Fan, Y. Feng, R. Song, Nonparametric independence screening in sparse ultra-high dimensional additive models, J. Amer. Statist. Assoc. 116 (2011) 544–557.
[9] P. Hall, On projection pursuit regression, Ann. Statist. 17 (1989) 573–588.
[10] W. Härdle, P. Hall, H. Ichimura, Optimal smoothing in single-index models, Ann. Statist. 21 (1993) 157–178.
[11] W. Härdle, T.M. Stoker, Investing smooth multiple regression by the method of average derivatives, J. Amer. Statist. Assoc. 84 (1989) 986–995.
[12] T.J. Hastie, R.J. Tibshirani, J. Friedman, The Elements of Statistical Learning, second ed., Springer, 2009.
[13] J. Horowitz, W. Härdle, Direct semiparametric estimation of a single-index model with discrete covariates, J. Amer. Statist. Assoc. 91 (1996) 1632–1640.
[14] M. Hristache, A. Juditsky, V. Spokoiny, Direct estimation of the index coefficient in a single-index model, Ann. Statist. 29 (2001) 595–623.
[15] J. Huang, J. Horowitz, F. Wei, Variable selection in nonparametric additive models, Ann. Statist. 38 (2010) 2282–2313.
[16] H. Ichimura, Semiparametric least squares (sls) and weighted sls estimation of single-index models, J. Econometrics 58 (1993) 71–120.
[17] H. Ichimura, S. Lee, Characterization of the asymptotic distribution of semiparametric m-estimators, J. Econometrics 159 (2010) 252–266.
[18] L. Li, X. Yin, Sliced inverse regression with regularizations, Biometrics 64 (2008) 124–131.
[19] L. Meier, S. van de Geer, P. Bühlmann, High-dimensional additive modeling, Ann. Statist. 37 (2009) 3779–3821.
[20] N. Meinshausen, Relaxed lasso, Comput. Statist. Data Anal. 52 (2007) 374–393.
[21] N. Meinshausen, B. Yu, Lasso-type recovery of sparse representations for high-dimensional data, Ann. Statist. 37 (2009) 246–270.
[22] S. Negahban, P. Ravikumar, M. Wainwright, B. Yu, A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers, Statist. Sci. 27 (2012) 538–557.
[23] L. Ni, R. Cook, C. Tsai, A note on shrinkage sliced inverse regression, Biometrika 92 (2005) 242–247.
[24] H. Peng, T. Huang, Penalized least squares for single index models, J. Statist. Plann. Inference 141 (2011) 1362–1379.
[25] D. Pollard, P. Radchenko, Nonlinear least-squares estimation, J. Multivariate Anal. 97 (2006) 548–562.
[26] P. Radchenko, G.M. James, Variable selection using adaptive nonlinear interaction structures in high dimensions, J. Amer. Statist. Assoc. 105 (2010) 1541–1553.
[27] P. Radchenko, G.M. James, Variable inclusion and shrinkage algorithms, J. Amer. Statist. Assoc. 103 (2008) 1304–1315.
[28] P. Ravikumar, J. Lafferty, H. Liu, L. Wasserman, Sparse additive models, J. R. Stat. Soc. Ser. B 71 (2009) 1009–1030.
[29] L. Schumaker, Spline Functions: Basic Theory, Wiley, New York, 1981.
[30] S. van de Geer, Empirical Processes in M-Estimation, Cambridge University Press, 1999.
[31] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes: With Applications to Statistics, Springer-Verlag, 1996.
[32] T. Wang, P.-R. Xu, L.-X. Zhu, Non-convex penalized estimation in high-dimensional models with single-index structure, J. Multivariate Anal. 109 (2012) 221–235.
[33] L. Wang, L. Yang, Spline estimation of single-index models, Statist. Sinica 19 (2009) 765–783.
[34] Q. Wang, X. Yin, A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse mave, Comput. Statist. Data Anal. 52 (2008) 4512–4512.
[35] Y. Wu, L. Li, Asymptotic properties of sufficient dimension reduction with a diverging number of predictors, Statist. Sinica 21 (2011) 707–730.
[36] Y. Xia, H. Tong, W. Li, L.-X. Zhu, An adaptive estimation of dimension reduction space (with discussion), J. R. Stat. Soc. Ser. B 64 (2002) 363–410.
[37] Y. Yu, D. Ruppert, Penalized spline estimation for partially linear single index models, J. Amer. Statist. Assoc. 97 (2002) 1042–1054.
[38] Z. Yu, L. Zhu, H. Peng, L. Zhu, Dimension reduction and predictor selection in semiparametric models, Biometrika 100 (2013) 641–654.
[39] J. Zhou, X. He, Dimension reduction based on constrained canonical correlation and variable filtering, Ann. Statist. 36 (2008) 1649–1668.

[40] S. Zhou, X. Shen, D. Wolfe, Local asymptotics for regression splines and confidence regions, Ann. Statist. 26 (1998) 1760–1782.
[41] L. Zhu, L. Li, R. Li, L. Zhu, Model-free feature screening for ultrahigh-dimensional data, J. Amer. Statist. Assoc. 106 (2011) 1464–1475.
[42] L. Zhu, L. Zhu, Nonconcave penalized inverse regression in single index models with high dimensional predictors, J. Multivariate Anal. 100 (2009) 862–875.