Econometric Theory, **0**, 2024, 1–43.
doi:10.1017/S0266466624000021

SEMIPARAMETRIC ESTIMATION AND VARIABLE SELECTION FOR SPARSE SINGLE INDEX MODELS IN INCREASING DIMENSION

CHAOHUA DONG
Zhongnan University of Economics and Law

YUNDONG TU
Peking University

This paper considers semiparametric sieve estimation in high-dimensional single index models. The use of Hermite polynomials in approximating the unknown link function provides a convenient framework to conduct both estimation and variable selection. The estimation of the index parameter is formulated from solutions obtained by the routine penalized weighted linear regression procedure, where the weights are used in order to tackle the unbounded support of the regressors. The resulting index parameter estimator is shown to be consistent and sparse, and the asymptotic normality for the estimators of both the index parameter and the link function is established. To perform variable selection in the ultra-high dimension case, we further suggest a forward regression screening method, which is shown to enjoy the sure independence screening property. This screening procedure can be used before the penalized variable selection to reduce the burden of dimensionality. Numerical results show that both the variable selection procedures and the associated estimators perform well in finite samples.

1. INTRODUCTION

The semiparametric single index model has been an important tool for practitioners to analyze the data with both linear and nonlinear features. By virtue of a linear index and a nonparametric link function, the model circumvents the curse of dimensionality but maintains the flexibility to capture possible nonlinear relationships. The estimation of the index parameter and the nonparametric function has

The authors thank the Editor, Co-Editor, and the two anonymous referees for constructive comments that helped improve the presentation of the paper. Dong thanks the partial support from the National Natural Science Foundation of China (Grant No. 72073143) and Fundamental Research Funds for the Central Universities, Zhongnan University of Economics and Law (Grant No. 2722022EG001). Tu acknowledges the partial support from the National Natural Science Foundation of China (Grant Nos. 72073002, 12026607, 92046021), the Center for Statistical Science at Peking University, and Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education. Address correspondence to Yundong Tu, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, China; e-mail: yundong.tu@gsm.pku.edu.cn. Address correspondence to Yungong Tu, Peking University; e-mail: yundong.tu@gsm.pku.edu.cn.

been intensively studied in the literature. This includes the semiparametric (profile) least square estimator (SLS, Ichimura, 1993; Hardle, Hall, and Ichimura, 1993), the maximum quasi-likelihood estimator (MQLE, Klein and Spady, 1993), the average derivative estimator (ADE, Hardle and Stocker, 1989; Power, Stock, and Stoker, 1989), the minimum average variance estimator (MAVE, Xia et al., 2002), the penalized spline estimator (PS, Yu and Ruppert, 2002; Ma, Liang, and Tsai, 2014), and the estimating function method (EFM, Cui, Hardle, and Zhu, 2011), to cite a few.

Nevertheless, the implementation of the above estimators often encounters practical challenges. For example, SLS, MQLE, PS, and EFM involve optimizations where there does not have an explicit solution. Consequently, numerical methods are inevitably required, for which the choice of the initial values tends to be critical and the convergence of the algorithm is often difficult to control, especially when the dimension of the index vector gets large. Therefore, these methods become less appealing for practitioners faced with high-dimensional covariates. Although ADE allows estimating the index parameter directly, it is found to suffer from the curse of dimensionality (Xia, 2006). In addition, MAVE may encounter the problem of data sparseness (Cui et al., 2011). See Cui et al. (2011) and Ma et al. (2014) for more discussions on the drawbacks of these methods.

This paper aims at addressing the above problems in a framework of high-dimensional semiparametric single-index models with sparsity. To achieve this goal, an explicit solution to the minimization of an objective function constructed from weighted least squares is proposed to facilitate both numerical and asymptotic analysis. In particular, an orthogonal series expansion of the nonparametric link function is firstly implemented in terms of Hermite polynomials, then the expansion of the single-index regression function is factorized into an additive form by the products of functions of unknown index vector and Hermite polynomials of regressors. Consequently, the semiparametric single index regression can be approximated by a linear parametric regression so that the estimators of the parameters can be obtained in explicit forms. This approach has been used in Dong, Gao, and Peng (2015), but that study suffers from the curse of dimensionality such that the number of covariates has to be less than four. It is our proposal of high-dimensional sparse single-index models that endows a promising usage to such an approach. Fortunately, through imposing penalization to a weighted least squares objective function, we are able to detect the model sparsity satisfactorily and establish the asymptotic properties of the proposed estimators.

There are several closely related papers, to cite a few, Ma et al. (2014), Peng and Huang (2011), and Radchenko (2015), in which the penalty functions have been used in the estimation of single index models. However, the current paper differs from the existing ones in at least two major aspects. First, the dimension of the index parameter in Ma et al. (2014) and Peng and Huang (2011) is assumed to be fixed and is not allowed to grow with the sample size n. Second, even though Radchenko (2015) allows the dimension of the index parameter to diverge with the sample size, the bounded support of the regressors has been imposed

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

like other papers. This is due to the fact that the spline approach used in these studies is restricted to functions defined on closed intervals for each n. The above limitations hamper the practical application of the single index model, especially in the presence of a number of potential regressors with unbounded support. By contrast, this paper allows the unboundedness of the regressors and combines the penalty function with a linear form approximation of the single index model simultaneously, from which the variable selection and estimation of the high-dimensional index vector and the link function are obtained by standard variable selection procedures such as SCAD of Fan and Li (2001). From these perspectives, our paper complements the above mentioned literature considerably.

From the practical point of view, when the dimension of the covariates is extremely high, the variable selection methods via the penalization approaches are known to suffer from drawbacks of computational inexpediency and algorithm instability (Fan, Samworth, and Wu, 2009). Therefore, such a variable selection problem in the single index model is beyond the scope of the existing studies, such as Radchenko (2015) mentioned earlier. The use of screening, as an alternative to penalized estimation, has proved successful in ultra-high-dimensional linear regressions and semiparametric/nonparametric regression setting, since the seminal work of Fan and Lv (2008). A number of screening procedures, including both model-dependent and model-free approaches, have been developed afterwards. See Kong, Xia, and Zhong (2019), Han (2019), Pan et al. (2019), Tu and Wang (2023), and the references therein for the recent advancements. In the single index setup, Gorst-Rasmussen and Scheike (2013) considered feature screening for survival data, and Zhang, Lian, and Yu (2020) considered variable selection in quantile regressions with the use of polynomial splines, thereby restricting the support of regressors to be bounded. In this study, we adapt the forward variable selection of Wang (2009) and Cheng, Honda, and Zhang (2016) to perform variable screening in the ultra-high-dimensional single index model, making use of the Hermite polynomial approximation as in the aforementioned penalized regression. This screening procedure, with the sure screening property that we shall demonstrate, can be used before the penalization-based variable selection approach to effectively reduce the dimensionality burden.

This paper contributes to the literature in the following aspects. First, we deal with regressors with possible unbounded support in the single index model using series expansion. In the sieve literature, researchers often restrict the support of the regressor to be compact (see, e.g., Newey, 1997; Ai and Chen, 2003; Peng and Huang, 2011; Ma et al., 2014; Belloni et al., 2015; Radchenko, 2015). This definitely hamstrings the use of plenty of regressors, especially the mostly encountered normal variables. Note that some recent studies (Chen and Christensen, 2015; Hansen, 2015) have paid effort to relax the restriction by adopting expanding intervals ($[-a_n, a_n]$ with $a_n \to +\infty$ depending on n) to approximate the real line. As an alternative, this paper allows the unbounded support for the regressor directly, and tackle the analytical challenges with the weighted estimation in terms of the density of the Hilbert space where the nonparametric function resides.

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

Second, we consider the sparsity in the index parameter, and also entertain the possibility that the link function is sparse as well. The dimensionality of the nonzero parameters (d_1) in the index vector and that of the index vector (d) are both allowed to grow with the sample size n. This setup is similar to the high-dimensional setting studied by Radchenko (2015). Further, while the number of basis functions used to approximate the unknown link function can diverge fast with the sample size, the vector of the coefficients in the combination can be sparse, that is, the notion of function sparsity raised by Belloni, Chernozhukov, and Wang (2014) in a nonparametric regression. To the best of our knowledge, such function sparsity has not been entertained in the semiparametric index setting so far. Moreover, the Hermite polynomial expansion used to obtain explicit solution leads to an entanglement among the index parameter and the parameters in the series approximation of the unknown function, which generates a complication in the dimensionality allowed in these parameters. This is detailed carefully in Assumption 3.5 in Section 3, which states that we may approximate the regression function by $K = \exp(n^a)$ terms with $0 < a < 1$. This feature coincides with the sparse linear regression literature where the number of regressors can be as large as $\exp(n^{\varepsilon})$ for some $0 < \varepsilon < 1$.

Third, we consider the recovery of the index parameter from its entanglement with the coefficients in the Hermite polynomial expansion of the nonparametric function. Dong et al. (2015) consider this recovery from the first block of the linear regression coefficient estimates. This paper shows that the recovery here is nonunique, and that the asymptotic variances of the resulted index parameter estimators are quite involved to compare. However, the major conclusion backs up the choice made in Dong et al. (2015) that the recovery from the first block is generally acceptable.

Fourth, asymptotic normality of the index parameter estimator and that of the refitted estimator for the link function are established. Under a set of regularity conditions, we are able to show that the penalized estimator in the linear approximation of the single index model is oracle and the penalization procedure achieves selection consistency. Furthermore, we show that the recovered index parameter estimator is $\sqrt{n}$-consistent and asymptotically normal. The resulted sieve estimator for the link function is shown to be consistent. However, the associated asymptotic distribution is quite involved to obtain. To circumvent this challenge, we alternatively consider the refitted estimator of the link function by plugging in the index estimator, the asymptotic normality of which is then attainable with the standard nonparametric convergence rate. The proposed estimation procedure is easily implementable with the standard algorithm used for linear regressions with a penalty, such as Fan and Li (2001). Simulation studies show that the estimation procedure performs quite well in finite samples.

Last but not least, we consider the more challenging variable selection problem where the dimension of regressors is ultra-high. Under this setting the traditional variable selection consistency becomes challenging or even impossible, while a number of screening procedures have been proved successful in linear regressions

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

and other nonparametric and semiparametric frameworks. To the best of our knowledge, this is the first work that studies the variable screening for the semiparametric single index regression model.[1] Based on the linear approximation used for variable selection, we show that a forward screening method married with the extended Bayesian information criterion-based stopping rule, adapted from Wang (2009) and Cheng et al. (2016), achieves the sure independence screening property. The method is found to be computationally expedient and demonstrates the sure screening property in finite sample experiments where the dimension of covariates is much larger than the sample size. Consequently, this screening method can precede the variable selection procedure when the latter fails to produce a reliable solution in the high-dimensional case.

The rest of the paper is organized as follows: Section 2 describes the sieve estimation approach for single index models, introduces the penalized estimation approach to simultaneously estimate the parameters and select relevant variables. Section 3 establishes the asymptotic properties of the penalized index estimator, and presents the estimator for the link function and its asymptotic normality. Section 4 presents a screening procedure to select variables in the ultra-high dimension, and proves its sure screening property. Numerical studies are presented in Section 5 to illustrate the finite sample performance of the proposed methods in both simulations and a real data example. The last section concludes with discussions on future research. All the proofs for the main results are collected in the Appendix, while the proofs for the auxiliary lemmas and some additional simulation results are relegated to Supplementary Material for space consideration.

Notation. $\|\cdot\|$ denotes Euclidean norm for vectors, or Frobenius norm for matrices; $\|\cdot\|_1$ is the ℓ_1 norm and $\|\cdot\|_\infty$ is the ℓ_∞ norm. Let $|A|$ stand for the cardinality of a set A. Given a vector $b \in \mathbb{R}^K$ and an index set $S \subset \{1,\ldots,K\}$, let b_S denote a vector in $\mathbb{R}^K$ whose jth element equals to b_j if $j \in S$, and zero otherwise, let $b_{(S)}$ denote the subvector of b that only selects the elements b_j's for $j \in S$, and let S^c be the complement of S. We use $\mathbf{0}$ to signify either a vector or matrix of zeros, whose dimension may be inferred from the context.

2. METHODOLOGY

2.1. The Model and Sieve Estimation

Consider the semiparametric single index model

$$y_t = g(\mathbf{x}_t'\boldsymbol{\theta}_0) + \epsilon_t, \quad t = 1,\ldots,n, \tag{2.1}$$

where g is an unknown link function, ϵ_t is an error sequence, $\mathbf{x}_t$ is a $d \times 1$ regressor. For identification purpose, the unknown index vector $\boldsymbol{\theta}_0$ satisfies $\|\boldsymbol{\theta}_0\| = 1$ with its

[1] Zhong et al. (2016) considered a screening problem in the conditional distribution, which is different from our regression model. They used the penalized quantile regression approach and adopted the distance correlation to perform screening (see Zhong et al., 2016 for details).

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

first element $\theta_{01} > 0$. We shall consider the case where the dimension d diverges with the sample size n.

Suppose that the regression function $g(z) \in L^2(\mathbb{R}, e^{-z^2/2}) := \{f(z) : \int f^2(z)e^{-z^2/2} dz < \infty\}$, a Hilbert space in which the inner product is defined by $\langle f_1,f_2\rangle = \int f_1(z)f_2(z)e^{-z^2/2}dz$ for $f_1(z),f_2(z) \in L^2(\mathbb{R}, e^{-z^2/2})$ and the induced norm $\|f\| = \sqrt{\langle f,f\rangle}$. Two functions $f_1(z)$ and $f_2(z)$ in $L^2(\mathbb{R}, e^{-z^2/2})$ are called orthogonal if $\langle f_1,f_2\rangle = 0$. Note that the tail of the density in the space $L^2(\mathbb{R}, e^{-z^2/2})$ is very thin such that the space is sufficiently large and contains at least all bounded functions, polynomials, power functions, and even some exponential functions.

It is worth pointing out that the supposition $g(z) \in L^2(\mathbb{R}, e^{-z^2/2})$ relaxes the sieve literature where the regression function is defined on a compact set (e.g., Newey, 1997; Ai and Chen, 2003; Ma et al., 2014; Belloni et al., 2015). The compactness definitely excludes some crucial regressors of interest such as normal and exponential distributions that are typically encountered in both theory and practice. We note that as long as one element of $\mathbf{x}_t$ possesses unbounded support the domain of $g(\cdot)$ should be treated as unbounded. To work with unbounded support, we will employ Hermite orthogonal polynomial sequence defined on the entire real line, and the unknown regression function will be estimated by the sieve method in the Hilbert space.

To begin, we introduce the following Hermite polynomial sequence $\{H_i(z)\}$ that forms an orthogonal basis of $L^2(\mathbb{R}, e^{-z^2/2})$. By definition,

$$H_i(z) = (-1)^i \exp(z^2/2)\frac{d^i}{dz^i}\exp(-z^2/2), \qquad i \geq 0, \tag{2.2}$$

are Hermite polynomials which are orthogonal, $\langle H_i, H_j\rangle = \sqrt{2\pi}\, i!\, \delta_{ij}$, where δ_{ij} is the Kronecker delta. Define $h_i(z) = (i!)^{-1/2}H_i(z)$ for the ease of exposition. Note that a crucial property about Hermite polynomials that shall be used frequently in our proofs is the uniform boundedness, that is, $\sup_{i\geq 0, z\in\mathbb{R}} |h_i(z)|e^{-z^2/4} < \infty$ (see Szego, 1975, p. 242).

Thus, any continuous function $g(z) \in L^2(\mathbb{R}, e^{-z^2/2})$ has an orthogonal series expansion in terms of $h_i(z)$, that is,

$$g(z) = \sum_{i=0}^{\infty} c_i h_i(z), \quad \text{where } c_i = \frac{1}{\sqrt{2\pi}}\langle g, h_i\rangle. \tag{2.3}$$

Let k be a positive integer, and define $g_k(z) = \sum_{i=0}^{k-1} c_i h_i(z)$ the truncation series and residue $\gamma_k(z) = \sum_{i=k}^{\infty} c_i h_i(z)$. It is known that, as $k \to \infty$, $g_k(z)$ converges to $g(z)$ in norm (i.e., $\|g_k - g\| \to 0$), whereas the pointwise convergence of $g_k(z)$ to $g(z)$ (i.e., $g_k(z) - g(z) \to 0$) on the real line relies on the smoothness of $g(z)$. With the above notations, model (2.1) can be written as

$$y_t = g_k(\mathbf{x}_t'\boldsymbol{\theta}_0) + \gamma_k(\mathbf{x}_t'\boldsymbol{\theta}_0) + \epsilon_t, \ \ t = 1, \ldots, n. \tag{2.4}$$

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

Meanwhile, by virtue of the property of the Hermite polynomials given by Lemma A.1 in the Appendix, we further write each term in $g_k(\mathbf{x}_t'\boldsymbol{\theta}_0)$ as

$$c_ih_i(\mathbf{x}_t'\boldsymbol{\theta}_0) = \sum_{|\mathbf{p}|=i} a_{i\mathbf{p}}(\boldsymbol{\theta}_0)\mathcal{H}_{\mathbf{p}}(\mathbf{x}_t), \;\; 0 \le i \le k-1, \tag{2.5}$$

where

$$\mathbf{p} = (p_1,\ldots,p_d)' \text{ with nonnegative integers } p_j, \;\; |\mathbf{p}| = p_1+\cdots+p_d,$$

$$a_{i\mathbf{p}}(\boldsymbol{\theta}_0) = \sqrt{\binom{i}{\mathbf{p}}}c_i\boldsymbol{\theta}_0^{\mathbf{p}}, \;\; \binom{i}{\mathbf{p}} = \frac{i!}{p_1!\ldots p_d!}, \;\; \boldsymbol{\theta}_0^{\mathbf{p}} = \prod_{j=1}^{d}\theta_{0j}^{p_j}, \quad \mathcal{H}_{\mathbf{p}}(\mathbf{x}_t) = \prod_{j=1}^{d} h_{p_j}(x_{t,j}). \tag{2.6}$$

Thus, each term $c_ih_i(\mathbf{x}_t'\boldsymbol{\theta}_0)$ corresponds to the sum of elements in the set $\{a_{i\mathbf{p}}(\boldsymbol{\theta}_0)\mathcal{H}_{\mathbf{p}}(\mathbf{x}_t), |\mathbf{p}| = i\}$. Note that if $c_i = 0$, each element in the set must be zero in view of the expression of $a_{i\mathbf{p}}(\boldsymbol{\theta}_0)$; if all terms in the set are zeros, we then conclude that $c_i = 0$ because at least $\theta_{01} > 0$. This fact will be utilized later to derive the estimators of nonzero c_i and $\boldsymbol{\theta}_0$ from the estimators of $a_{i\mathbf{p}}(\boldsymbol{\theta}_0)$.

To write all the terms in $g_k(\mathbf{x}_t'\boldsymbol{\theta}_0) = \sum_{i=0}^{k-1}\sum_{|\mathbf{p}|=i} a_{i\mathbf{p}}(\boldsymbol{\theta}_0)\mathcal{H}_{\mathbf{p}}(\mathbf{x}_t)$ in order, we introduce an ordering relationship among all $\mathbf{p}$ such that $|\mathbf{p}| = i$ for $i \le k-1$.

DEFINITION 2.1. *Let $\mathcal{P}_i = \{\mathbf{p} = (p_1,\ldots,p_d) : |\mathbf{p}| = i\}$, where i is a nonnegative integer. For any $\hat{\mathbf{p}}, \check{\mathbf{p}} \in \mathcal{P}_i$, we say $\hat{\mathbf{p}} = (\hat{p}_1,\ldots,\hat{p}_d) < \check{\mathbf{p}} = (\check{p}_1,\ldots,\check{p}_d)$ if there exists an ℓ $(1 < \ell \le d)$ such that $\hat{p}_j = \check{p}_j$ for all $j = 1,\ldots,\ell-1$ but $\hat{p}_\ell < \check{p}_\ell$.*

Noting that $|\mathcal{P}_i| = \binom{i+d-1}{d-1}$, the total number of the terms in $g_k(\mathbf{x}_t'\boldsymbol{\theta}_0)$ is $K = \sum_{i=0}^{k-1}\binom{i+d-1}{d-1} = \binom{k+d-1}{d} = O(k^d)$. In view of the expansion in (2.4) and (2.5) and the ordering introduced above, we may write model (2.1) in matrix form as

$$\mathbf{Y} = Z\boldsymbol{\beta}_0 + \boldsymbol{\gamma} + \mathbf{e}, \tag{2.7}$$

where $\mathbf{Y} = (y_1,\ldots,y_n)'$, $Z = (Z_k(\mathbf{x}_1),\ldots,Z_k(\mathbf{x}_n))'$ an $n \times K$ matrix, in which $Z_k(\mathbf{x}_t)$ is a column vector of dimension K consisting of all terms $\mathcal{H}_{\mathbf{p}}(\mathbf{x}_t)$ for all $\mathbf{p} : |\mathbf{p}| = i$ and $i = 0,1,\ldots,k-1$ in ascending order of i and $\mathbf{p}$ according to Definition 2.1, $\boldsymbol{\beta}_0 = (\beta_{01},\ldots,\beta_{0K})'$ consists of all coefficients $a_{i\mathbf{p}}(\boldsymbol{\theta}_0)$ in the same order as the elements in $Z_k(\cdot)$, $\boldsymbol{\gamma} = (\gamma_k(\mathbf{x}_1'\boldsymbol{\theta}_0),\ldots,\gamma_k(\mathbf{x}_n'\boldsymbol{\theta}_0))'$, $\mathbf{e} = (\epsilon_1,\ldots,\epsilon_n)'$.

We remark that β_{0j} and $a_{ip}(\boldsymbol{\theta}_0)$ are mutually determined uniquely. First, given $a_{i\mathbf{p}}(\boldsymbol{\theta}_0)$ with $0 \le i \le k-1$ and $\mathbf{p} \in \mathcal{P}_i$, if $i = 0$ (implying $\mathbf{p} = (0,\ldots,0)$ only), then $j = 1$, that is, $\beta_{01} = a_{0\mathbf{p}}(\boldsymbol{\theta}_0)$; if $i \ge 1$, then $j = \sum_{\ell=0}^{i-1}|\mathcal{P}_\ell| + u_{\mathbf{p}}$, where $u_{\mathbf{p}}$ is a positive integer such that $\mathbf{p}$ is the $u_{\mathbf{p}}$th element in $\mathcal{P}_i$. Second, given j with $1 \le j \le K$, if $\sum_{\ell=0}^{i}|\mathcal{P}_\ell| \le j < \sum_{\ell=0}^{i+1}|\mathcal{P}_\ell|$ for some $0 \le i \le k-1$, then $\beta_{0j} \mapsto a_{i\mathbf{p}}(\boldsymbol{\theta}_0)$, where $|\mathbf{p}| = i$ and $\mathbf{p}$ is the $u_{\mathbf{p}}$th element in $\mathcal{P}_i$ with $u_{\mathbf{p}} = j - \sum_{\ell=0}^{i-1}|\mathcal{P}_\ell|$ (the convention $\sum_{\ell=0}^{-1} = 0$ applies if any). In addition, β_{0j} and c_i are also mutually determined uniquely via $\text{Sub}(\boldsymbol{\beta}_0, i) = (a_{i\mathbf{p}}(\boldsymbol{\theta}_0), |\mathbf{p}| = i)'$, the subvector of $\boldsymbol{\beta}_0$ associated with c_i, for $i = 0,\ldots,k-1$. Consequently, both the index parameter $\boldsymbol{\theta}_0$ and the sieve coefficients, $c_0, c_1, \ldots, c_{k-1}$, can then be recovered from $\boldsymbol{\beta}_0$. The details will be discussed in Section 2.3 when we derive their estimators.

Note that $\boldsymbol{\beta}_0$ is typically estimated by the ordinary least squares, whenever $K < n$. However, it is worth noting that the optimal sieve order to approximate univariate unknown functions that minimizes the mean squared error loss is $k = O(n^{1/5})$. Such a choice of k would make the number of parameters $K = O(k^d)$ in (2.7) comparable to the sample size n if d is 5 or larger, which makes the ordinary least square estimation of parameters in (2.7) infeasible. Below we shall consider the case where $\boldsymbol{\theta}_0$ is sparse and contains (a large number of) zeros, and potentially the link function $g(\cdot)$ is sparse in the sense of Belloni et al. (2014). In this case, the penalized estimation techniques can lend support to the estimation of $\boldsymbol{\beta}_0$, and subsequent inferences on both $\boldsymbol{\theta}_0$ and $g(\cdot)$.

2.2. Sparsity and Penalized Estimation

It is interesting to note that the sparsity of $\boldsymbol{\beta}_0$ could stem from two sources. The main source is the sparsity of $\boldsymbol{\theta}_0$. To be precise, decompose $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are of dimensions d_1 and d_2, respectively, with $\boldsymbol{\theta}_2 = \mathbf{0}$ and $d = d_1 + d_2$. Conformably, decompose $\mathbf{p} = (\mathbf{p}_1', \mathbf{p}_2')'$. Then, a typical subvector of $\boldsymbol{\beta}_0$, $a_{i\mathbf{p}}(\boldsymbol{\theta}_0) = \sqrt{\binom{i}{\mathbf{p}}} c_i \boldsymbol{\theta}_1^{\mathbf{p}_1} \boldsymbol{\theta}_2^{\mathbf{p}_2}$ is zero if $\mathbf{p}_2$ is nonzero; because d_2 is large due to the sparsity of $\boldsymbol{\theta}_0$, $\boldsymbol{\beta}_0$ is sparse. The other source might be the sparsity of the link function, that is, there are considerable number of coefficients $c_i = 0$ in the expansion of $g(z)$. This happens, for example, when $g(z)$ belongs to some finite-dimensional subspace of the L^2 space, so $c_i = 0$ for all $i \geq k_0$ with some fixed k_0 (Belloni et al., 2014). As a result, $\mathrm{Sub}(\boldsymbol{\beta}_0, i) = \mathbf{0}$ for all $i \geq k_0$, leading to further sparsity in $\boldsymbol{\beta}_0$. Noting that the latter sparsity could depend on the type of sieve basis used, only the former type of sparsity is required for the theory developed below. The sparse feature of $\boldsymbol{\beta}_0$ permits the effective identification of the nonzero components, based on which variable selection can be achieved.

Before presenting the penalized estimation, we emphasize that most studies from the existing sieve literature (e.g., Newey, 1997; Ai and Chen, 2003; Belloni et al., 2015; Chen and Christensen, 2015, etc.) require the support of regressors to be bounded. Nonetheless, we allow some elements of $\mathbf{x}_t$ possessing unbounded support on $\mathbb{R}$, which makes the practical choice of explanatory variables much broader. The direct consequence is that the support of the unknown function $g(\cdot)$ will become unbounded. Inevitably, this relaxation gives rise to an enormous challenge for the asymptotic analysis when Hermite polynomials are utilized. This is because the sieve method normally requires that the vector of basis functions of dimension k be order of $O(\sqrt{k})$ or $O(k)$ in Euclidean norm, uniformly over the support of the regressor, in order to achieve the asymptotic normality of nonparametric sieve estimator (see Newey, 1997; Chen and Shen, 1998; Belloni et al., 2015, among others). Fortunately, we find that weighted estimation with the density of the Hilbert space can eschew the unboundedness of the associated norm (i.e., $\|Z_k(\mathbf{x})\|$ in this paper), which renders the asymptotic theory applicable. To do so, denote by $\widetilde{\mathbf{x}}_t$ a subvector of $\mathbf{x}_t$, whose support might be unbounded and dimension $\widetilde{d}$ is fixed. It is noteworthy that one is not required to know which elements have unbounded support, instead it suffices to know the subvector complement to $\widetilde{\mathbf{x}}_t$ that has bounded support.

To proceed, define $W = \mathrm{diag}(w(\widetilde{\mathbf{x}}_1), \ldots, w(\widetilde{\mathbf{x}}_n))$, where $w(\widetilde{\mathbf{x}}_t) = \exp(-\|\widetilde{\mathbf{x}}_t)\|^2/2)$ for $t = 1, \ldots, n$. The penalized weighted least squares estimator of $\boldsymbol{\beta}_0$ is defined as

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^K} Q_n(\boldsymbol{\beta}) \equiv L_n(\boldsymbol{\beta}) + \sum_{j=1}^{K} P_n(|\beta_j|), \tag{2.8}$$

where $L_n(\boldsymbol{\beta}) := \frac{1}{n}(\mathbf{Y} - Z\boldsymbol{\beta})'W(\mathbf{Y} - Z\boldsymbol{\beta})$ and $P_n(\cdot)$ is a penalty function, such as the SCAD of Fan and Li (2001), to be discussed in the next section. Denote $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_K)'$ for later use. The use of the weighting of the function space will turn the norm requirement on $Z_k(\mathbf{x}_t)$ to be that on $Z_k(\mathbf{x}_t)w^{1/2}(\widetilde{\mathbf{x}}_t)$ for the asymptotic analysis, which is of order $O(\sqrt{K})$ uniformly in $\mathbf{x}_t$. In this way, we shall show that the unbounded support issue is overcome.

2.3. Recovery of the Index Vector and Link Function

Of our primary interest are the index vector $\boldsymbol{\theta}_0$ and the link function $g(\cdot)$. In what follows, we shall discuss the relationship among $\boldsymbol{\theta}_0$, $g(\cdot)$ and $\boldsymbol{\beta}_0$, based on which the estimators will be constructed.

For $i = 0$, we simply let $\widehat{c}_0 = \widehat{\beta}_1$ because $c_0 = \beta_{01}$. For $i = 1, \ldots, k-1$, let $Q_i = (\mathbf{0}, I_d, \mathbf{0})_{d\times K}$, where the first zero matrix has dimension $d \times [(d+i-1)!/d!(i-1)!]$, while the second zero matrix has conformable dimension according to $\boldsymbol{\beta}_0$. Then Q_i facilitates to pick up the first d elements from the block $\mathrm{Sub}(\boldsymbol{\beta}_0; i)$. That is,

$$Q_i\boldsymbol{\beta}_0 = (c_i\theta_{01}^i, \sqrt{i}c_i\theta_{01}^{i-1}\theta_{02}, \ldots, \sqrt{i}c_i\theta_{01}^{i-1}\theta_{0d})',$$

which implies that

$$\boldsymbol{\theta}_0 = \frac{1}{c_i\theta_{01}^{i-1}} D_i Q_i \boldsymbol{\beta}_0, \quad \text{where } D_i := \mathrm{diag}(1, 1/\sqrt{i}, \ldots, 1/\sqrt{i}), \tag{2.9}$$

if $c_i \neq 0$.

Furthermore, we take out the elements in $\mathrm{Sub}(\boldsymbol{\beta}_0; i)$ that have the form of ith power. They are

$$\beta_{0,i_1} = c_i\theta_{01}^i, \ldots, \beta_{0,i_d} = c_i\theta_{0d}^i, \tag{2.10}$$

where the corresponding subindexes are $i_1 = \frac{(d+i-1)!}{d!(i-1)!} + 1$, $i_2 = i_1 + \frac{(d+i-1)!}{(d-1)!i!} - \frac{(d+i-2)!}{(d-2)!i!}$, $i_3 = i_2 + \frac{(d+i-2)!}{(d-2)!i!} - \frac{(d+i-3)!}{(d-3)!i!}$, …, $i_d = \frac{(d+i)!}{d!i!}$. The relationship in (2.10) suggests that

$$c_i = \mathrm{sgn}(\beta_{0,i_1}) \left(\sum_{j=1}^{d} \left(\beta_{0,i_j}^2 \right)^{1/i} \right)^{i/2}, \tag{2.11}$$

since $\theta_{01} > 0$ by identification.

By (2.10) and (2.11), we can estimate c_i by

$$\widehat{c}_i = \mathrm{sgn}(\widehat{\beta}_{i_1}) \left(\sum_{j=1}^{d} (\widehat{\beta}_{i_j}^2)^{1/i} \right)^{i/2},$$

and estimate θ_{01} by

$$\widehat{\theta}_{01} = (\widehat{\beta}_{i_1}/\hat{c}_i)^{1/i},$$

provided that $\widehat{c}_i \neq 0$. In conjunction with (2.9), we can estimate $\boldsymbol{\theta}_0$ by

$$\widehat{\boldsymbol{\theta}} = \frac{1}{\widehat{c}_i \widehat{\theta}_{01}^{i-1}} D_i Q_i \widehat{\boldsymbol{\beta}}, \tag{2.12}$$

provided that $\widehat{c}_i \widehat{\theta}_{01} \neq 0$.

Finally, let $\widehat{J} = \{j : \ \widehat{\theta}_{0j} \neq 0\}$ be the estimator for the true index set $J = \{j : \theta_{0j} \neq 0\}$, and $\widehat{I} = \{i : \ \widehat{c}_i \neq 0, 0 \leq i \leq k-1\}$ be that for $I = \{i : \ c_i \neq 0, 0 \leq i \leq k-1\}$. After obtaining $\widehat{c}_i, i = 0, \ldots, k-1$, one can estimate the link function by the plug-in estimator $\widehat{g}(z) := \sum_{i \in \widehat{I}} \widehat{c}_i h_i(z)$. However, its asymptotic properties are quite involved due to the nonlinear relationship between $\widehat{c}_i$ and $\widehat{\boldsymbol{\beta}}$.

We circumvent the above challenge via a refit of the nonparametric function in the following after obtaining the estimate of the index vector. Note that the estimate $\widehat{\boldsymbol{\beta}}$ suggests that the unknown function $g(z)$ should have the form $\sum_{i \in \widehat{I}} c_i h_i(z)$, indicating that the sparsity of the link function can be achieved. Given $\widehat{\boldsymbol{\theta}}$ and $\widehat{I}$, we then estimate $c_{\widehat{I}} = (c_i, i \in \widehat{I})$ from the model

$$y_t = \sum_{i \in \widehat{I}} c_i h_i(\widehat{\boldsymbol{\theta}}' \mathbf{x}_t) + \gamma_k(\widehat{\boldsymbol{\theta}}' \mathbf{x}_t) + e_t, \ t = 1, \ldots, n. \tag{2.13}$$

The weighted least squares estimator is defined as

$$\widetilde{c}_{\widehat{I}} = \underset{c}{\mathrm{argmin}} \sum_{t=1}^{n} \left(y_t - \sum_{i \in \widehat{I}} c_i h_i(\widehat{\boldsymbol{\theta}}' \mathbf{x}_t) \right)^2 w(\widehat{\boldsymbol{\theta}}' \bar{\mathbf{x}}_t), \tag{2.14}$$

with the explicit form $\widetilde{c}_{\widehat{I}} = (\widehat{U}'\widehat{W}\widehat{U})^{-1}\widehat{U}'\widehat{W}\mathbf{Y}$, where $\widehat{U} = (\Phi_{(\widehat{I})}(\widehat{\boldsymbol{\theta}}'\mathbf{x}_1), \ldots, \Phi_{(\widehat{I})}(\widehat{\boldsymbol{\theta}}'\mathbf{x}_n))'$ an $n \times |\widehat{I}|$ matrix, $\widehat{W} = \mathrm{diag}(w(\widehat{\boldsymbol{\theta}}'\bar{\mathbf{x}}_1), \ldots, w(\widehat{\boldsymbol{\theta}}'\bar{\mathbf{x}}_n))$, $w(z) = \exp(-z^2/2)$, $\Phi_{(\widehat{I})}(z)$ is the subvector of $\Phi(z)$ with the subscript contained in $\widehat{I}$, and $\bar{\mathbf{x}}_t$ is the d-vector that replaces the elements of $\mathbf{x}_t$ with bounded support (i.e., the complement of $\widetilde{\mathbf{x}}_t$) by zeros. Note that the weight constructed above only involves the potentially unbounded subvector. Then, we may define $\widetilde{g}(z) = \Phi_{(\widehat{I})}(z)'\widetilde{c}_{\widehat{I}}$ for any $z \in \mathbb{R}$. Note that the above explicit expression for $\widetilde{c}_{\widehat{I}}$ facilitates the derivation of the asymptotic normality of $\widetilde{g}(z)$. This complements the literature such as Belloni et al. (2014) where only the approximation order is obtained.

3. ASYMPTOTIC THEORY

3.1. Assumptions

Before presenting the technical assumptions, we introduce some notations. Let $S_0 = \text{supp}(\boldsymbol{\beta}_0) = \{j : 1 \leq j \leq K, |\beta_{0j}| > 0\}$ be the index set for the true model, and let $s_0 = |S_0|$ be the cardinality of S_0.

Assumption 3.1. (a) $\{\epsilon_t, \mathbf{x}_t\}$ is an independent and identically distributed sequence drawn from $(\epsilon, \mathbf{x})$, and $E(\epsilon|\mathbf{x}) = 0$ almost surely (a.s.), $E(\epsilon^2|\mathbf{x}) = \sigma_e^2$ a.s. and $E(\epsilon^4|\mathbf{x}) = \mu_4 < \infty$ a.s.; (b) there exist $b_1, b_2 > 0$, such that for any $u > 0$, we have $P(|\epsilon| > u) \leq \exp(-(u/b_1)^{b_2})$.

Assumption 3.2. $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^d$, where Θ is a convex and compact set and $\boldsymbol{\theta}_0$ is an interior point of Θ. For each $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\theta}'\mathbf{x}_t$ has density $f_{\boldsymbol{\theta}}(v)$ such that $\sup_{\boldsymbol{\theta} \in \Theta} f_{\boldsymbol{\theta}}(v) \leq C\exp(-v^2/2)$ for all large $|v|$ and some constant $C > 0$.

Assumption 3.3. The link function $g(z) \in L^2(\mathbb{R}, e^{-z^2/2})$ is differentiable up to ν-th order on $\mathbb{R}$ and $g^{(\nu)}(z) \in L^2(\mathbb{R}, e^{-z^2/2})$.

The above conditions are often used in the linear parametric and semiparametric context. Assumption 3.1 requires that the regressor and the error be uncorrelated and the error sequence be conditionally homoscedastic. The conditional homoscedasticity may be restrictive for some applications, but serves to provide reasonable approximation when the data are properly transformed (e.g., by log transformation). The exponential tail condition imposed in the assumption is satisfied by normal random variables and other variables that have compact support. See Assumption 4.3 of Fan and Liao (2014) and A4 of Radchenko (2015, p. 277). This assumption is used to confine the score function in the asymptotic development.

Assumption 3.2 is commonly used for the parametric space and the single-index model context (e.g., Dong et al., 2015). Here, $\boldsymbol{\theta}'\mathbf{x}_t$ is stipulated to have a tail not fatter than a normal distribution that covers a variety of variables, especially normal variables and those that have compact support. There are many research papers that exclude normal variables by imposing the compactness of its support. As noted in the Introduction, the recent literature starts to relax the restriction by adopting expanding intervals to approximate the whole real line. By contrast, taking advantage of the density in the function space, we are able to tackle the unbounded support in a quite natural and easy way.

Assumption 3.3 imposes a smoothness order for the link function that expedites the convergence of the orthogonal series expansion. See Lemma A.2 in the

Appendix. The concrete requirement on the order ν of differentiability is implied by Lemma A.2 and Assumption 3.5 implicitly. Note that smoothness order ν in Assumption 3.3 affects the quantity s_0, and roughly speaking, $s_0 = O(k^{d_1})$ for finite ν, while $s_0 = O(k_0^{d_1})$ for some fixed k_0 when $\nu = \infty$, implied from Lemma A.2. Because $d_1 \ll d$, it is clear that $s_0 \ll K$ where $K = O(k^d)$.

Suppose that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ where $\boldsymbol{\beta}_2 = \mathbf{0}$. Define the score function $F_n(\boldsymbol{\beta}) = -\frac{2}{n}Z'W(\mathbf{Y} - Z\boldsymbol{\beta})$ that is the partial derivative of $L_n(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Then, $F_n(\boldsymbol{\beta}_0) = -\frac{2}{n}Z'W(\boldsymbol{\gamma} + \mathbf{e})$. Further, let the Hessian matrix be $H_n(\boldsymbol{\beta}) = \frac{2}{n}Z'WZ$ which is independent of $\boldsymbol{\beta}$. Normally, when $K > n$, the smallest eigenvalue of the Hessian matrix would be zero that violates the identifiability.

For any $S \subset \{1, \ldots, K\}$ with $s = |S|$, denote by $Z_{(S)}$ the matrix eliminating all jth column, $j \notin S$, of Z, so that its dimension is $n \times s$; $\boldsymbol{\beta}_{(S)}$ the short vector of $\boldsymbol{\beta}_S$ removing all zeros at $j \notin S$, so its dimension is s. Remember that $\boldsymbol{\beta}_S \in \mathbb{R}^K$ whereas $\boldsymbol{\beta}_{(S)} \in \mathbb{R}^s$. Hence, $Z_{(S)}\boldsymbol{\beta}_{(S)} = Z\boldsymbol{\beta}_S$ for any $\boldsymbol{\beta}$ and S. Now, for any given S, define $F_{n(S)}(\boldsymbol{\beta}_{(S)}) = -\frac{2}{n}Z_{(S)}'W(\mathbf{Y} - Z_{(S)}\boldsymbol{\beta}_{(S)})$ as the derivative vector of $L_n(\boldsymbol{\beta}_S)$ with respect to all β_j, for $j \in S$. Similarly, define $H_{n(S)}(\boldsymbol{\beta}_{(S)}) = \frac{2}{n}Z_{(S)}'WZ_{(S)}$, the second derivative matrix of $L_n(\boldsymbol{\beta}_S)$ with respect to all β_j, for $j \in S$, the dimension of which is $s \times s$.

Suppose that $P_n(\cdot)$ belongs to the class of folded concave penalty functions in Fan and Li (2001). For any $\mathbf{v} = (v_1, \ldots, v_{s_0})' \in \mathbb{R}^{s_0}$ with $v_j \neq 0$, $\forall j$, define

$$\phi(\mathbf{v}) = \limsup_{\epsilon \to 0+} \max_{j \le s_0} \sup_{(u_1, u_2) \subset O(|v_j|, \epsilon)} -\frac{P_n'(u_2) - P_n'(u_1)}{u_2 - u_1},$$

where $O(\cdot, \cdot)$ is the neighborhood with specified center and radius, respectively. This implies that $\phi(\mathbf{v}) = \max_{j \le s_0} -P_n''(|v_j|)$ if P_n'' is continuous. Also, for the true parameter $\boldsymbol{\beta}_0$, let

$$\zeta_n = \frac{1}{2}\min\{|\beta_{0j}| : \ \beta_{0j} \neq 0, j = 1, \ldots, K\},$$

represent the strength of the signal.

The following assumptions related to the penalty function and the signal strength are needed for the main results.

Assumption 3.4. The penalty function $P_n(u)$ satisfies (i) $P_n(0) = 0$; (ii) $P_n(u)$ is concave, nondecreasing on $[0, \infty)$, and has a continuous derivative $P_n'(u)$ for $u > 0$; $(iii)\sqrt{s_0}P_n'(\zeta_n) = o(\zeta_n)$; (iv) There exists $c > 0$ such that $\sup_{\mathbf{v} \in O(\boldsymbol{\beta}_1, c\zeta_n)} \phi(\mathbf{v}) = o(1)$.

This assumption is mostly encountered in the literature, and is satisfied by many penalty functions with proper choice of tuning parameters, such as the L_q penalty with $q \le 1$, hard-thresholding (Antoniadis, 1996), SCAD (Fan and Li, 2001), and MCP (Zhang, 2010). We shall discuss the conditions on ζ_n after the following assumption.

Assumption 3.5. Suppose that:

(i) $\sup_{\|\boldsymbol{\beta}_{S_0}-\boldsymbol{\beta}_0\|\leq\zeta_n/4}\phi(\boldsymbol{\beta}_{(S_0)}) = o((s_0\log(K))^{-1/2})$, $P'_n(\zeta_n) = o(1/\sqrt{s_0 n})$, $\sqrt{s_0\log(K)/n} = o(\zeta_n)$;

(ii) $\sqrt{s_0}P'_n(\zeta_n)+\sqrt{s_0\log(K)/n}+\sqrt{s_0}\|\gamma_k(z)\| = o(P'_n(0^+))$;

(iii) $\|\gamma_k(z)\| = o(1/\sqrt{s_0 n})$;

(iv) $H_{n(S_0)}(\boldsymbol{\beta}_{(S_0)})$ *has eigenvalues bounded below from zero and above from infinity uniformly in* n.

Assumption 3.5(i) imposes further requirements on the penalty function, the strength of the minimal signal and the support of the true parameter. For the SCAD or MCP penalty function with tuning parameter $\lambda_n = o(\zeta_n)$, we have $P'_n(\zeta_n) = 0$, $\sup_{\|\boldsymbol{\beta}_{S_0}-\boldsymbol{\beta}_0\|\leq\zeta_n/4}\phi(\boldsymbol{\beta}_{(S_0)}) = 0$ and $P'_n(0^+) = \lambda_n$. The three conditions in Assumption 3.5(i), together with Assumption 3.4(iii), are therefore fulfilled. The condition 3.5(ii) implies that K can be as large as $\exp(n^{\varepsilon})$ for some $0 < \varepsilon < 1$. Noting that $K = O(k^d)$, one possibility for the choice of k and d is $k = \exp(n^a)$ and $d = n^b$ with $a, b > 0, a+b = \varepsilon$. The rate $k = \exp(n^a)$ is comparable with that in Belloni et al. (2014) where the sparsity of function is studied in a nonparametric setting.

Another possibility to fulfill Assumption 3.5(i) is $\zeta_n = a\lambda_n - \widetilde{\zeta}_n$ where $\sqrt{s_0}\widetilde{\zeta}_n = o(\zeta_n)$. Along with the SCAD penalty, this implies $\sqrt{s_0}P'_n(\zeta_n) = \sqrt{s_0}\widetilde{\zeta}_n/(a-1) = o(\zeta_n)$, so Assumption 3.4(iii) is satisfied. Moreover, all conditions in Assumption 3.5 related to ζ_n are satisfied. Specifically, Assumption 3.5(i) that requires $P'_n(\zeta_n) = o(1/\sqrt{s_0 n})$ and $\sqrt{s_0(\log K)/n} = o(\zeta_n)$ is fulfilled as long as $\widetilde{\zeta}_n = o(1/\sqrt{s_0 n})$ and $\sqrt{s_0(\log K)/n} = o(\lambda_n)$, since $P'_n(\zeta_n) \sim \widetilde{\zeta}_n$ and $\zeta_n \sim \lambda_n$. In addition, Assumption 3.5(ii) that requires $\sqrt{s_0}P'_n(\zeta_n) = o(\lambda_n)$ is readily valid because of Assumption 3.5(i) $\sqrt{s_0}P'_n(\zeta_n) = o(\zeta_n)$ and $\zeta_n \sim \lambda_n$. To summarize, the condition on ζ_n allows it to converge to zero either at the same rate or slower than λ_n. See Assumptions 4.1, 4.5, and 4.6 in Fan and Liao (2014) for more detailed discussion on similar assumptions in the linear regression model.

Assumption 3.5(iii) is an undersmoothing condition often used in sieve estimation to eliminate effect of the truncated residue, though it seems strong at appearance. To meet this requirement, the regression function usually has to be very smooth because k cannot diverge fast. However, this condition is easily satisfied, with a suitable smoothness order and fast divergence of k, in the current context where $\mathbf{c} = (c_0, \ldots, c_{k-1})'$ is allowed to possess sparsity (Belloni et al., 2014).

Assumption 3.5(iv) is commonly imposed for the Hessian matrix in the literature, and is equivalent to the sparse Riesz condition in Zhang and Huang (2008, p. 1572). For a similar requirement, see Condition A.2 in Belloni et al. (2015, p. 347), Assumption 4 in Chen and Christensen (2015, p. 450), Assumption 3 in Donald, Imbens, and Newey (2009, p. 31), Assumption 3.2 in Ai and Chen (2007, p. 14), Assumption 3.2 in Ai and Chen (2003, p. 1803), Assumption 2 in Newey (1997, p. 149). While the structure of the Hessian matrix in our study is different from those in the aforementioned papers, this condition is easily fulfilled if one is willing to assume that, for example, d_1 is fixed and $g(\cdot)$ is sparse, implying that $s_0 = |S_0|$ is fixed.

3.2. Asymptotic Properties

Recall that $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ with $\boldsymbol{\theta}_2 = \mathbf{0}$, and the dimensions of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are d_1 and d_2, respectively, with $d = d_1 + d_2$. To formulate $\boldsymbol{\theta}_1$ in terms of S_0, the support of $\boldsymbol{\beta}_0$, we need to introduce selection matrices that withdraw subvectors from $\boldsymbol{\beta}_1$. Define the matrix $R_i = (\mathbf{0}, I_{|J|}, \mathbf{0})_{|J|\times s_0}$, where $|J| = d_1, |S_0| = s_0$ and the first zero matrix has $(d_1 + i - 1)/d_1!\,(i-1)!$ columns, and the second zero matrix has conformable columns. Then, $R_i\boldsymbol{\beta}_1$ is the vector consisting of the first $|J|$ elements of $\boldsymbol{\beta}_1$ associated with c_i, and has elements of the form either $c_i\theta_{01}^i$ or $\sqrt{i}c_i\theta_{01}^{i-1}\theta_{0j}$ for some $j \in \{2, \ldots, d\}$. Hence, it is easy to see that $\boldsymbol{\theta}_1 = \frac{1}{c_i\theta_{01}^{i-1}}A_iR_i\boldsymbol{\beta}_1$, provided $c_i \neq 0$, with the $|J|$-dimensional diagonal matrix $A_i = \mathrm{diag}(1, 1/\sqrt{i}, \ldots, 1/\sqrt{i})$. We thus consider $\widehat{\boldsymbol{\theta}}_1^{(i)} = \frac{1}{\hat{c}_i\hat{\theta}_{01}^{i-1}}A_iR_i\widehat{\boldsymbol{\beta}}_1$ as the estimate of $\boldsymbol{\theta}_1$ derived from the subvector $\mathrm{Sub}(\widehat{\boldsymbol{\beta}}; i)$, provided that $\widehat{c}_i \neq 0$, with S_0 and J replaced by their estimates $\widehat{S}$ and $\widehat{J}$, respectively.

Let Ω and Ψ be the probability limits of $Z_{(S_0)}'WZ_{(S_0)}/n$ and $Z_{(S_0)}'W^2Z_{(S_0)}/n$, respectively, the convergence of which are given in Lemma A.3. In addition, we assume that all eigenvalues of Ω and Ψ are bounded below from zero and above from infinity.

THEOREM 3.1. *Let Assumptions 3.1–3.5 hold.*

1. *We have $P(\widehat{J} = J) \to 1$ as $n \to \infty$.*
2. *Let $c_i \neq 0$, $1 \leq i \leq k-1$. For any $\boldsymbol{\alpha} \in \mathbb{R}^{|d_1|}$ with $\|\boldsymbol{\alpha}\| = 1$, as $n \to \infty$,*

$$\sqrt{n}\sigma_{ni}^{-1}\frac{\widehat{\beta}_{i_1}}{\widehat{\theta}_{01}}\boldsymbol{\alpha}'(\widehat{\boldsymbol{\theta}}_1^{(i)} - \boldsymbol{\theta}_1) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\sigma_{ni}^2 = \boldsymbol{\alpha}'B_{ni}\Omega^{-1}\Psi\Omega^{-1}B_{ni}'\boldsymbol{\alpha}\sigma_e^2$, $B_{ni} := \left[I_{d_1} + \frac{1}{\theta_{01}}\boldsymbol{\theta}_1\ell_1'\right]^{-1}\left[A_iR_i + \frac{1}{\theta_{01}}\boldsymbol{\theta}_1\ell_{i_1}'\right]$, $i_1 = (d+i-1)!/d!\,(i-1)!$, ℓ_{i_1} is an s_0-vector whose i_1th element is 1 and elsewhere zero, while ℓ_1 is a d_1-vector whose first element is 1 and elsewhere zero.

Theorem 3.1 establishes the consistency of the index set estimator $\widehat{J}$, and the consistency and the asymptotic normality of the estimator $\widehat{\boldsymbol{\theta}}_1^{(i)}$. Nevertheless, the variance formula σ_{ni}^2 is a bit complicated. This is due to the entanglement of $\boldsymbol{\theta}_0$ and c_i in $\mathrm{Sub}(\beta_0; i)$, that is, $\beta_{0j} \mapsto c_i\sqrt{\binom{i}{\mathbf{p}}}\boldsymbol{\theta}_0^{\mathbf{p}}$, and we derive the estimates of c_i and $\boldsymbol{\theta}_0$ from that of $\mathrm{Sub}(\boldsymbol{\beta}_0; i)$. Given fixed i, the convergence rate becomes root-n, which is comparable with existing results, such as those in Theorem 2 of Chang, Chen, and Chen (2015, p. 288). Note that in B_{ni} the matrix $I_{d_1} + \frac{1}{\theta_{01}}\boldsymbol{\theta}_1\ell_1'$ is invertible, since

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

it is lower triangular with $(2, 1, \ldots, 1)$ being the diagonal elements. The inverse can be obtained easily because on the lower triangular part only the first column may be nonzero.

As noted above, we might have more than one estimator of $\boldsymbol{\theta}_1$, $\widehat{\boldsymbol{\theta}}_1^{(i)}$, derived from the ith block $\mathrm{Sub}(\boldsymbol{\beta}_0; i)$ where $c_i \neq 0$, for $i \leq k-1$. We recommend, in terms of the efficiency, the one derived from the subvector $\mathrm{Sub}(\beta_0; i)$, where $c_i \neq 0$ and i is the smallest number among all possible blocks.[2] The reason is as follows. The variance of $\sqrt{n}\boldsymbol{\alpha}'(\widehat{\boldsymbol{\theta}}_1^{(i)} - \boldsymbol{\theta}_1)$ is approximately $\sigma_{ni}^2\theta_{01}^2/\beta_{0i_1}^2 = \sigma_{ni}^2/c_i^2\theta_{01}^{2(i-1)}$. Noting first that in the factor matrix $A_iR_i + \frac{1}{\theta_{01}}\boldsymbol{\theta}_1\ell_{i_1}'$ (in σ_{ni}^2), when i is sufficiently large, $A_i = \mathrm{diag}(1, 1/\sqrt{i}, \ldots, 1/\sqrt{i})$ is close to the matrix where only the left-top element is one and elsewhere is zero. On the other hand, noting the facts that $0 < \theta_{01} \leq 1$ due to the identification condition, and $c_i \to 0$ when $i \to \infty$ by the Parseval equality in Hilbert space, in general, the smaller the i is, the smaller the variance. This observation also backs up the choice made in Dong et al. (2015), where d has been assumed to be fixed.

To present the asymptotic distribution for $\widetilde{g}(z)$, define $U = (\Phi_{(\widehat{I})}(\boldsymbol{\theta}_0'\mathbf{x}_1), \ldots, \Phi_{(\widehat{I})}(\boldsymbol{\theta}_0'\mathbf{x}_n))'$ and $W_0 = \mathrm{diag}(w(\boldsymbol{\theta}_0'\bar{\mathbf{x}}_1), \ldots, w(\boldsymbol{\theta}_0'\bar{\mathbf{x}}_n))$, where $\bar{\mathbf{x}}_t$ is the d-vector that replaces the elements of $\mathbf{x}_t$ with bounded support by zeros. Furthermore, let Ξ be the probability limit of $\frac{1}{n}U'W_0U$, whose (i,j) element is $E[h_i(\boldsymbol{\theta}_0'\mathbf{x}_t)h_j(\boldsymbol{\theta}_0'\mathbf{x}_t)w(\boldsymbol{\theta}_0'\bar{\mathbf{x}}_t)]$, and Σ be that of $\frac{1}{n}U'W_0^2U$, whose (i,j) element is $E[h_i(\boldsymbol{\theta}_0'\mathbf{x}_t)h_j(\boldsymbol{\theta}_0'\mathbf{x}_t)w^2(\boldsymbol{\theta}_0'\bar{\mathbf{x}}_t)]$ for $i,j \in \widehat{I}$.

THEOREM 3.2. *In addition to Assumptions 3.1–3.5, suppose that* $d_1^{3/2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = o_P(1)$, *and that all eigenvalues of* Ξ *and* Σ *are uniformly bounded away from zero and above from infinity.*

1. *We have* $P(\widehat{I} = I) \to 1$ *as* $n \to \infty$.
2. *For any* $z \in \mathbb{R}$, *if* $\sqrt{n/d_1}|\gamma_k(z)| = o(1)$,

$$\sigma_n^{-1}\frac{\sqrt{n}}{\|\Phi_{(\widehat{I})}(z)\|}[\widetilde{g}(z) - g(z)] \xrightarrow{d} \mathcal{N}(0,1), \tag{3.1}$$

as $n \to \infty$, *where* $\sigma_n^2 = \overline{\Phi}_{(\widehat{I})}(z)'\Xi^{-1}\Sigma\,\Xi^{-1}\overline{\Phi}_{\widehat{I}}(z)\sigma_e^2$ *with* $\overline{\Phi}_{(\widehat{I})}(z) = \Phi_{(\widehat{I})}(z)/\|\Phi_{(\widehat{I})}(z)\|$.

The conditions on the eigenvalues are typically required in the derivation of asymptotic normality. Meanwhile, the undersmoothing condition on the residue $\gamma_k(\cdot)$ is also mostly encountered in the literature, to eliminate the effect of truncation error (see Comment 4.3 in Belloni et al., 2015, p. 352). This can be fulfilled if the function g is sparse and smooth with certain order, as explained earlier below

[2] Alternatively, it is also possible to derive an estimator of $\boldsymbol{\theta}_1$ from a set of blocks $\mathrm{Sub}(\boldsymbol{\beta}_0; i)$, where $c_i \neq 0$, $i \leq k-1$. It could be potentially more efficient as more information could be employed. However, such an estimator is more complicated for implementation and its theoretical property is also quite involved to explore. We shall report findings along this line elsewhere.

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

Assumption 3.5. On the other hand, the condition $d_1^{3/2}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\| = o_P(1)$ puts a bit more restriction on the dimension d_1, which however is not difficult to satisfy. In addition, the convergence rate of $\widetilde{g}(z)$ is $\sqrt{n/|\widehat{I}|}$ in view of $|\widehat{I}|$ being the number of basis functions used in the refitted model. This is fairly standard in the sieve literature, such as Newey (1997) and Ai and Chen (2003), among others.

4. SCREENING IN ULTRA-HIGH DIMENSION

Turning to the scenario when the number of regressors is much larger than the sample size, the variable selection methods via the penalization approaches in (2.8) are known to suffer from drawbacks of computational inexpediency and algorithm instability, as pointed out by Fan et al. (2009), among others. Since the pioneer work of Fan and Lv (2008), it is well-known that screening can effectively reduce the dimensionality by removing irrelevant regressors. This section proposes to use the forward screening approach, adapted from Wang (2009) and Cheng et al. (2016), to select variables in the single index setting.

The idea of screening works as follows. We start with B_1 that is either an empty set or a set that contains some "must-have" regressors, for example, those implied from related economic theory. Then we need consider sequentially whether we should add some $\ell \in B_1^c = \{1,\dots,d\}\setminus B_1$ into B_1 to form $B_2 = B_1 \cup \{\ell\}$ and go to the next step, or we should stop searching for any additional covariate if certain stopping criterion is met.

To state the screening procedure for the single index model in a general way, suppose that we have already selected a set B of important indices, followed by considering whether we should add some $\ell \in B^c$ into B to obtain a new augmented set of important indices. Let $\mathbf{x}_{t,B}$ be the subvector of $\mathbf{x}_t$ that consists of all elements $x_{t,i}$ of $\mathbf{x}_t$, where $i \in B$. Consider similarly a single index model

$$y_t = g_B(\mathbf{x}_{t,B}'\boldsymbol{\theta}_B) + \epsilon_{t,B}, \quad t = 1,\dots,n, \tag{4.1}$$

where the link function $g_B(z)$ and $\boldsymbol{\theta}_B$ together minimize $\mathbb{E}[y_t - g(\mathbf{x}_{t,B}'\theta)]^2$ over $g \in L^2(\mathbb{R}, e^{-z^2/2})$ and $\|\theta\| = 1$. Thus, similar to (2.4) and (2.5), we have

$$y_t = \sum_{i=0}^{k-1} c_{i,B} h_i(\mathbf{x}_{t,B}'\boldsymbol{\theta}_B) + \gamma_{k,B}(\mathbf{x}_{t,B}'\boldsymbol{\theta}_B) + \epsilon_{t,B}, \quad t = 1,\dots,n, \tag{4.2}$$

$$c_{i,B} h_i(\mathbf{x}_{t,B}'\boldsymbol{\theta}_B) = \sum_{|\mathbf{p}|=i} a_{i\mathbf{p}}(\boldsymbol{\theta}_B)\mathcal{H}_{\mathbf{p}}(\mathbf{x}_{t,B}), \quad 0 \le i \le k-1, \tag{4.3}$$

where $\mathbf{p}$'s are multiple indices with dimension $|B|$, and

$$a_{i\mathbf{p}}(\boldsymbol{\theta}_B) = \sqrt{\binom{i}{\mathbf{p}}} c_{i,B}\boldsymbol{\theta}_B^{\mathbf{p}}, \quad \boldsymbol{\theta}_B^{\mathbf{p}} = \prod_{j=1}^{|B|} \theta_{B,j}^{p_j}, \quad \mathcal{H}_{\mathbf{p}}(\mathbf{x}_{t,B}) = \prod_{j=1}^{|B|} h_{p_j}(x_{t,B,j}). \tag{4.4}$$

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

Note that all summands in $c_{i,B}h_i(\mathbf{x}'_{t,B}\boldsymbol{\theta}_B)$ in equation (4.3) form a set $\{a_{i\mathbf{p}}(\boldsymbol{\theta}_B)\mathcal{H}_{\mathbf{p}}(\mathbf{x}_{t,B}), |\mathbf{p}|=i\}$ that has cardinality $\binom{i+|B|-1}{|B|-1}$, where $\mathcal{H}_{\mathbf{p}}(\mathbf{x}_{t,B})$ contains known (basis) functions evaluated at the observations $\mathbf{x}_{t,B}$. With the above notations, model (4.1) is written as

$$y_t = Z_k(\mathbf{x}_{t,B})'\boldsymbol{\beta}_B + \gamma_{k,B}(\mathbf{x}'_{t,B}\boldsymbol{\theta}_B) + \epsilon_{t,B}, \quad t=1,\ldots,n, \tag{4.5}$$

where $Z_k(\mathbf{x}_{t,B})$ is a column vector of dimension $K_B = \sum_{i=0}^{k-1}\binom{i+|B|-1}{|B|-1} = \binom{k+|B|-1}{|B|}$ consisting of all terms $\mathcal{H}_{\mathbf{p}}(\mathbf{x}_{t,B})$ for all $\mathbf{p}:|\mathbf{p}|=i$ and $i=0,1,\ldots,k-1$ in some order of i and $\mathbf{p}$, $\boldsymbol{\beta}_B = (\beta_{B,1},\ldots,\beta_{B,K_B})'$ stands for the vector stacking all unknown coefficients $a_{i\mathbf{p}}(\boldsymbol{\theta}_B)$ in the same order as the elements in $Z_k(\cdot)$. In matrix form, the equation (4.5) can be written as

$$\mathbf{Y} = Z_B\boldsymbol{\beta}_B + \boldsymbol{\gamma}_B + \mathbf{e}_B, \tag{4.6}$$

where $\mathbf{Y}=(y_1,\ldots,y_n)'$, $Z_B = (Z_k(\mathbf{x}_{1,B}),\ldots,Z_k(\mathbf{x}_{n,B}))'$ is an $n\times K_B$ matrix, $\mathbf{e}_B = (\epsilon_{1,B},\ldots,\epsilon_{n,B})'$, and $\boldsymbol{\gamma}_B = (\gamma_k(\mathbf{x}'_{1,B}\boldsymbol{\theta}_B),\ldots,\gamma_k(\mathbf{x}'_{n,B}\boldsymbol{\theta}_B))'$.

We estimate $\boldsymbol{\beta}_B$ by the ordinary least squares, that is,

$$\widehat{\boldsymbol{\beta}}_B = \underset{\boldsymbol{\beta}\in\mathbb{R}^{K_B}}{\text{argmin}}\,\|\mathbf{Y}-Z_B\boldsymbol{\beta}\|^2, \tag{4.7}$$

which gives $\widehat{\boldsymbol{\beta}}_B = (Z'_BZ_B)^{-1}Z'_B\mathbf{Y}$. Then, we have $\widehat{\mathbf{e}}_B = \mathbf{Y}-Z_B\widehat{\boldsymbol{\beta}}_B$, the estimated residues for model (4.1).

Now we consider whether there is an index from B^c that we should add to B to obtain a new set of important variables, or stop screening. Denote $B(\ell)=B\cup\{\ell\}$ for $\ell\in B^c$. Let $\mathbf{x}_{t,B(\ell)}$ be the subvector of $\mathbf{x}_t$ defined similarly as $\mathbf{x}_{t,B}$. Consider the extended single index model for $B(\ell)$, $\ell\in B^c$,

$$y_t = g_{B(\ell)}(\mathbf{x}'_{t,B(\ell)}\boldsymbol{\theta}_{B(\ell)}) + \epsilon_{t,B(\ell)}, \quad t=1,\ldots,n, \tag{4.8}$$

where $\|\boldsymbol{\theta}_{B(\ell)}\|=1$.

Following the same procedure as (4.1)–(4.7), we can have the estimate of residue $\widehat{\mathbf{e}}_{B(\ell)} = \mathbf{Y}-Z_{B(\ell)}\widehat{\boldsymbol{\beta}}_{B(\ell)}$ for model (4.8). Indeed, we can write for $t=1,\ldots,n$,

$$y_t = Z_k(\mathbf{x}_{t,B(\ell)})\boldsymbol{\beta}_{B(\ell)} + \boldsymbol{\gamma}_{k,B(\ell)}(\mathbf{x}'_{t,B(\ell)}\boldsymbol{\theta}_{B(\ell)}) + \mathbf{e}_{t,B(\ell)}, \tag{4.9}$$

where the $K_{B(\ell)}$-dimensional vector $Z_k(\mathbf{x}_{t,B(\ell)})$ contains elements $\mathcal{H}_{\mathbf{p}}(x_{t,B(\ell)})$. Here, the multiple indices $\mathbf{p}=(p_1,\ldots,p_{|B|+1})$ satisfy $|\mathbf{p}|=i$ for $i=0,\ldots,k-1$. Further, if $p_{|B|+1}=0$, the term $\mathcal{H}_p(x_{t,B(\ell)})$ is free of the newly added covariate $x_{t,S(\ell),|B|+1}$, since $h_0\equiv 1$. Thus, we can separate the vector $Z_k(\mathbf{x}_{t,B(\ell)})$ into two subvectors. The first is exactly $Z_k(\mathbf{x}_{t,B})$ in (4.5), while the second, denoted by $\tilde{Z}_k(\mathbf{x}_{t,B(\ell)})$, consists of all $\mathcal{H}_{\mathbf{p}}(\mathbf{x}_{t,B(\ell)})$ where $p_{|B|+1}\neq 0$. Thus, equation (4.9) can be written in matrix form as

$$\mathbf{Y} = Z_B\boldsymbol{\beta}_B + \tilde{Z}_{B(\ell)}\boldsymbol{\beta}_\ell + \boldsymbol{\gamma}_{B(\ell)} + \mathbf{e}_{B(\ell)}. \tag{4.10}$$

Define $H_B = I_n - Z_B(Z_B'Z_B)^{-1}Z_B$ and

$$\widehat{\sigma}_B^2 = \frac{1}{n}\|\widehat{\mathbf{e}}_B\|^2, \text{ and } \widehat{\sigma}_{B(\ell)}^2 = \frac{1}{n}\|\widehat{\mathbf{e}}_{B(\ell)})\|^2. \tag{4.11}$$

We then have

$$n\widehat{\sigma}_B^2 - n\widehat{\sigma}_{B(\ell)}^2 = \|\widehat{e}_B\|^2 - \|\widehat{e}_{B(\ell)}\|^2 = \widehat{\boldsymbol{\beta}}_\ell'(\tilde{Z}_{B(\ell)}' H_B \tilde{Z}_{B(\ell)})\widehat{\boldsymbol{\beta}}_\ell = n\,\mathbb{E}\|\breve{Z}_{B(\ell)}\boldsymbol{\beta}_\ell\|^2(1+o_P(1)),$$

where $\breve{Z}_{B(\ell)}$ is the residue of projecting $\tilde{Z}_{B(\ell)}$ onto the space spanned by Z_B. If $\ell \in J$ the set of indices of all important variables (i.e., $\boldsymbol{\beta}_\ell \neq \mathbf{0}$), the quantity above would be sufficiently larger than that corresponding to $\ell \in J^c$. Therefore, we choose ℓ^* such that $\widehat{\sigma}_{B(\ell^*)}^2 = \min_{\ell \in B^c} \widehat{\sigma}_{B(\ell)}^2$ as the candidate index, in which we have high confidence that $\ell^* \in J \setminus B$ provided it is not empty.

To determine whether or not to add the index ℓ^* to the set B of selected indices, we apply the extended Bayesian information criterion (EBIC), adapted from Wang (2009) and Cheng et al. (2016),

$$\text{EBIC}(B) = n\log(\widehat{\sigma}_B^2) + K_B(\log(n) + 2\eta\log(d)), \tag{4.12}$$

where the constant $\eta \geq 0$, and when $\eta = 0$ the EBIC reduces to BIC. We should include ℓ^* into B, if EBIC$(B(\ell^*))<$ EBIC(B); if the EBIC increases we should stop the search. The screening procedure is summarized as below.

The Screening Procedure:

Initial step	Start with B_1 that is either an empty set or a set that contains some "must-have" regressors, e.g., those implied from related economic theory. Compute EBIC(B_1).
Sequential step	In the $(m+1)$th step, compute $\widehat{\sigma}_{B_m(\ell)}^2$ for all $\ell \in B_m^c$, and find $\ell_{m+1}^* = \underset{\ell \in B_m^c}{\text{argmin}}\, \widehat{\sigma}_{B_m(\ell)}^2.$ Then, let $B_{m+1} = B_m \cup \{\ell_{m+1}^*\}$ and compute EBIC(B_{m+1}).
Stopping rule	Stop and declare B_m to be the set of selected covariate indexes if EBIC$(B_{m+1}) >$ EBIC(B_m); otherwise, change m to $m+1$ in the sequential step and continue searching for the next candidate regressor.

As noted in Cheng et al. (2016), the forward screening scheme combined with the EBIC or BIC stopping rule may stop a little too early due to rounding errors, in which case not all relevant variables are selected. This could happen, for example, when the stopping criterion value drops for one step, then increases in the next, and drops again. To enhance its practical performance, the forward selection process can continue until the stopping criterion value continuously increases for several (e.g., three) consecutive steps before stopping.

We next turn to establish the sure independence property of the above forward regression screening procedure. The following condition further specifies Assumption 3.1(b).

Assumption 4.1. There is a positive constant C_ϵ such that for any given $u \in \mathbb{R}$,

$$\mathbb{E}(\exp(u\epsilon)|\mathbf{x}) \le \exp(C_\epsilon u^2/2).$$

This sub-Gaussian assumption has been widely used in the literature for variable selection and sure screening, such as Wang (2009) and Cheng et al. (2016). It allows for normal random variables or variables that have bounded support.

Assumption 4.2. For any $M \ge d_1$, there exist positive constants $c(M)$ and $C(M)$ (possibly dependent on M), and $\mu > 0$, such that

$$c(M)k^{-\mu} \le \lambda_{\min}(E[\tilde{Z}_k(\mathbf{x}_{1,B})\tilde{Z}_k(\mathbf{x}_{1,B})']) \le \lambda_{\max}(E[\tilde{Z}_k(\mathbf{x}_{1,B})\tilde{Z}_k(\mathbf{x}_{1,B})']) \le C(M)k^{\mu}$$

uniformly in B with $|B| < M$, where $\tilde{Z}_k(\mathbf{x}_{1,B})$ is defined by (4.10).

This condition describes the eigenvalues of regression matrix. Here, we allow the minimum eigenvalue decaying to zero and the maximum eigenvalue diverging to infinity at certain rates, although in some ideal situation these eigenvalues are bounded from both below and above. See, for example, Proposition 2.1 of Belloni et al. (2015).

Assumption 4.3. Suppose that as $n \to \infty$,

$$(a)\ \ \frac{k^{\nu-3\mu}}{d_1^2} \to \infty; \ \ (b)\ \ \frac{n(\log(n))^{-\tau}}{d_1^2 k^{M+3\mu}} \to \infty; \ \ (c)\ \ M\log(d) = O(K_B(\log(n))^{\tau}),$$

for some $0 < \tau < 1$ and all $|B| \le M$, where ν and μ are specified by Assumptions 3.3 and 4.2, respectively.

Note that Assumption 3.3 underlies the above assumption to ensure a quick convergence of the orthogonal expansion for the link function. Combining Assumption 4.1 and Lemma C.1 of Dong et al. (2016), this yields $\|\gamma_k(\cdot)\|_{L^2} = o(k^{-\nu})$. As a result, the truncation error does not affect the procedure of the screening asymptotically. The conditions in Assumption 4.3 specify technical requirements on the divergence rates of eigenvalues of the signal matrices and the truncation parameter, and the divergence rates for the number of candidate covariates. Indeed, if we stipulate $\log(d) = n^{\varepsilon}$ and $k = [n^{\kappa}]$ for some $\varepsilon, \kappa > 0$, Assumption 4.3(c) will be fulfilled when $Mn^{\varepsilon} = O(n^{M\kappa}(\log(n))^{\tau})$ or $Mn^{-(M\kappa-\varepsilon)} = O((\log(n))^{\tau})$, as a consequence of taking $M = O((\log(n))^{\tau})$.

The following theorem determines the lower bound of the reduction in the sum of squared residuals when $J \not\subset B$.

THEOREM 4.1. *Suppose that Assumptions 3.1–3.3 and 4.1–4.3 hold. For all subsets $B \subset \{1,\dots,d\}$ with $|B \cup J| \le M$ and $J \not\subset B$, where J is the index set of $\boldsymbol{\theta}_1$, we have*

$$\max_{\ell \in B^c}[n\widehat{\sigma}_B^2 - n\widehat{\sigma}_{B(\ell)}^2] \geq \frac{n\, c_0^4\, c^2(M)}{d_1^2\, \|g\|^2\, C(M)\, k^{3\mu}},$$

with probability tending to one as $n \to \infty$, provided that $c_0 \neq 0$.

Note that $c_0 = \int g(x)e^{-x^2}dx$ and it is assumed that $c_0 \neq 0$ in the above theorem. In case that $c_0 = 0$, the above result continues to hold if we replace c_0 by the first c_i such that $c_i = \int g(x)h_i(x)e^{-x^2}dx \neq 0$ for $i \leq k-1$, as can be seen from the proof.

For notational simplicity, denote $D_M \equiv \frac{c_0^4\, c^2(M)}{d_1^2\, \|g\|^2\, C(M)\, k^{3\mu}}$. Let T_M be the smallest integer greater than or equal to $\mathrm{Var}(y)/D_M$. Following from Theorem 4.1, Corollary 4.1 gives a sufficient condition for screening consistency of the forward selection procedure.

COROLLARY 4.1. *In addition to the conditions in Theorem 4.1, if $T_M \leq M - d_1$ and $k^M(\log(n) + \eta \log(d)) = o(nD_M)$, then we have $J \subset B_m$ for some $m \leq T_M$ with probability tending to one.*

COROLLARY 4.2. *In addition to the conditions in Theorem 4.1, if $J \not\subset B_{m-1}$, but $J \subset B_m$ and $m \leq M$, then the forward screening procedure stops at the mth step with probability tending to one.*

With the aid of Theorem 4.1, the above corollaries further show that, under certain regularity conditions, the screening procedure is consistent, and is expected to stop at the mth step with $J \subset B_m$. These results are relevant in both theory and applications. Post the screening procedure, one could then use the methodology of penalized estimation stated in Section 2 to find exactly the true set of relevant regressors J, and then estimate the index vector and the unknown link function. The associated asymptotic properties as demonstrated earlier in Section 3 continue to hold, the detailed discussions of which are omitted due to space consideration.

5. NUMERICAL RESULTS

5.1. Simulations

In this section, we investigate the finite sample performance of the proposed variable selection procedure in semiparametric single index models. We consider the following four data generating processes (DGPs):

$$\begin{aligned}
&DGP \quad 1: \qquad y_t = \frac{\exp(\mathbf{x}_t'\boldsymbol{\theta}_0)}{1+\exp(\mathbf{x}_t'\boldsymbol{\theta}_0)} + e_t;\\
&DGP \quad 2: \qquad y_t = \Phi(\mathbf{x}_t'\boldsymbol{\theta}_0) + e_t;\\
&DGP \quad 3: \qquad y_t = \mathbf{x}_t'\boldsymbol{\theta}_0 + (\mathbf{x}_t'\boldsymbol{\theta}_0)^2 + e_t;\\
&DGP \quad 4: \qquad y_t = \exp(\mathbf{x}_t'\boldsymbol{\theta}_0) + e_t,
\end{aligned}$$

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

where $\mathbf{x}_t$ is $d \times 1$ normal random vector with zero mean and identity covariance matrix, $\boldsymbol{\theta}_0 = (\mathbf{1}'_{d_1}, \mathbf{0}'_{d_2})'/\sqrt{d_1}$ with $\mathbf{1}_{d_1}$ denoting the $d_1 \times 1$ vector of ones and $\mathbf{0}_{d_2}$ the $d_2 \times 1$ vector of zeros, the positive integers d_1, d_2 satisfy $d_1 + d_2 = d$, e_t is standard normal and generated independently from $\mathbf{x}_t$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The regression functions in DGP 1 and DGP 2 are bounded, while those in DGP 3 and DGP 4 are not.

5.1.1. *Variable Screening.* We first evaluate the proposed variable screening methods via the information criteria. Specifically, we compare the finite sample performance of the Akaike information criterion (AIC), Bayesian information criterion (BIC), and EBIC. At the initial step of the forward selection, we let $B_1 = \{1\}$. To prevent the forward procedures from stopping too early and missing some true variables, we terminated the sequential selection only if the stopping criterion value increases for three consecutive steps, as noted in Section 4. The value of the parameter η in the definition of EBIC is taken as $\eta = 1 - \log(n)/(3\log(d))$, following Chen and Chen (2008), and Cheng et al. (2016). Note that BIC can be regarded as a special version of EBIC when $\eta = 0$, while AIC can be obtained from BIC via replacing $\log(n)$ with 2. This shows that the penalty terms of AIC, BIC, and EBIC are getting larger in turns and hence the model selected by AIC is the largest, followed by the one selected by BIC, while the one selected by EBIC is the smallest.

To do so, we consider the aforementioned DGP 1–4 and generate the data accordingly for $n = 100, 200, 400$ and $d = 100, 400, 800$. The true number of covariates is set to be $d_1 = 2$, with the index vector to be $(1/\sqrt{2}, 1/\sqrt{2})'$. The residual standard deviation is set as $\sigma = 0.2$. To save space, we only report the results for the truncation parameter $k = 3$. As we find that the AIC method tends to select too many variables than necessary, we stop the screening procedure if the number of regressors selected reaches 10. We find that such a stopping rule does not affect the performance of BIC and EBIC. In Table 1, we report the average numbers of true positive (TP) and false positive (FP) selections, and their standard deviations in parentheses for the three methods. It is recognized that the larger the TP values and the smaller the FP values are, the better the associated approach performs.

Several findings are in order from Table 1. First, the forward screening using the AIC always selects the maximal number of covariates allowed and is able to select the true important covariates. This shows that the penalization used in AIC is not large enough to lead to desired sparsity. Second, both the BIC and EBIC are seen to select fewer covariates and are able to detect all the important covariates. Generally, their performance tends to improve as the sample size increases, but becomes worse as the total dimension of covariates d increases. Third, the EBIC tends to select fewer covariates than BIC does. Although the TP for EBIC is slightly smaller than that for BIC, especially when the sample size is small, that of the former quickly catches up to that of the latter as sample size increases to 200. Noticeably, the FP of the former is much smaller than that of the latter, indicating

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

Table 1. Average numbers of true positive (TP) and false positive (FP) over 1,000 repetitions and their robust standard deviations (in parentheses) for the AIC, BIC, and EBIC methods under DGPs 1–4.

			$n = 100$		$n = 200$		$n = 400$	
DGP	d	Method	TP	FP	TP	FP	TP	FP
1	400	AIC	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)
		BIC	2.00(0.00)	3.88(2.87)	2.00(0.00)	1.29(0.48)	2.00(0.00)	1.11(0.31)
		EBIC	2.00(0.07)	0.74(0.44)	2.00(0.00)	1.00(0.00)	2.00(0.00)	1.00(0.00)
	800	AIC	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)
		BIC	2.00(0.00)	6.38(2.66)	2.00(0.00)	1.57(0.61)	2.00(0.00)	1.18(0.39)
		EBIC	1.98(0.14)	0.51(0.50)	2.00(0.00)	1.00(0.00)	2.00(0.00)	1.00(0.00)
2	400	AIC	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)
		BIC	2.00(0.00)	3.67(2.87)	2.00(0.00)	1.25(0.46)	2.00(0.00)	1.07(0.26)
		EBIC	1.98(0.13)	0.59(0.49)	2.00(0.00)	1.00(0.00)	2.00(0.00)	1.00(0.00)
	800	AIC	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)
		BIC	2.00(0.00)	6.18(2.79)	2.00(0.00)	1.46(0.59)	2.00(0.00)	1.15(0.37)
		EBIC	1.92(0.27)	0.32(0.47)	2.00(0.00)	1.00(0.00)	2.00(0.00)	1.00(0.00)
3	400	AIC	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)
		BIC	2.00(0.00)	7.87(0.81)	2.00(0.00)	4.18(1.50)	2.00(0.00)	3.72(1.24)
		EBIC	1.98(0.13)	0.29(0.47)	2.00(0.00)	0.51(0.59)	2.00(0.00)	0.62(0.65)
	800	AIC	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)
		BIC	2.00(0.00)	8.00(0.00)	2.00(0.00)	5.14(1.75)	2.00(0.00)	4.61(1.50)
		EBIC	1.94(0.23)	0.31(0.47)	2.00(0.00)	0.50(0.56)	2.00(0.00)	0.69(0.64)
4	400	AIC	1.99(0.15)	8.00(0.00)	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)
		BIC	1.99(0.15)	7.96(0.54)	2.00(0.00)	5.78(1.97)	2.00(0.00)	5.92(1.92)
		EBIC	1.91(0.35)	0.64(0.69)	2.00(0.04)	1.05(0.85)	2.00(0.00)	1.52(1.14)
	800	AIC	1.99(0.15)	8.00(0.00)	2.00(0.00)	8.00(0.00)	2.00(0.00)	8.00(0.00)
		BIC	1.99(0.15)	8.00(0.00)	2.00(0.00)	6.50(1.78)	2.00(0.00)	6.76(1.68)
		EBIC	1.86(0.41)	0.57(0.66)	2.00(0.05)	1.02(0.85)	2.00(0.00)	1.63(1.08)

that EBIC is more efficient in reducing the dimensionality than BIC. Overall, all three methods are able to detect the important variables, which demonstrates that the sure screening property is possessed even in finite samples.

5.1.2. *Variable Selection.* To evaluate the performance of the variable selection procedure, we introduce several popular measures that are adopted in, for example, Ma et al. (2014). Let S be any candidate model and S_0 be the true model. We say that the model S is overfitted, correctly fitted, and underfitted if $S_0 \subset S$ (but $S_0 \neq S$), $S_0 = S$, and $S_0 \not\subseteq S$, respectively. We calculate the percentage of models overfitted (OF), correctly fitted (CF), and underfitted (UF) in 1,000 replications

Table 2. Variable selection results for DGP 1 and DGP 2, $d = 5$.

		Unweighted					Weighted				
n	k	C	IC	OF	CF	UF	C	IC	OF	CF	UF
						DGP 1					
200	3	96.6	11.5	27.9	65.3	6.8	97.5	24.9	51.3	43.8	4.9
	5	95.5	9.5	23.1	67.9	9.0	96.3	23.2	48.0	44.6	7.4
	7	94.3	9.2	20.5	68.0	11.5	94.4	21.5	45.0	43.8	11.2
800	3	100.0	0.6	1.7	98.3	0.0	100.0	3.6	10.6	89.4	0.0
	5	100.0	0.3	0.9	99.1	0.0	100.0	3.7	11.0	89.0	0.0
	7	100.0	0.3	0.8	99.2	0.0	100.0	4.0	10.9	89.1	0.0
						DGP 2					
200	3	97.5	13.0	31.7	63.2	5.0	97.7	25.6	54.3	41.1	4.6
	5	95.7	10.4	25.0	66.4	8.6	97.2	21.6	46.2	48.1	5.7
	7	95.0	10.8	24.9	64.7	10.1	95.3	20.5	42.9	47.7	9.2
800	3	100.0	0.5	1.5	98.5	0.0	100.0	4.0	11.7	88.3	0.0
	5	100.0	0.2	0.5	99.5	0.0	99.9	3.5	9.5	90.3	0.2
	7	100.0	0.3	0.9	99.1	0.0	100.0	3.9	11.3	88.7	0.0

for each DGP with $n = 100, 200, 400, 800$. We also compute the average number of nonzero elements in θ_0 that are correctly (C) estimated to be nonzero (normalized by d_1), the average number of zero elements that are incorrectly (IC) estimated to be nonzero (normalized by d_2). Due to space limit, only some selected results are reported in Tables 2–5 for $d = 5, 10$ and $d_1 = 2$. The results are calculated with the sieve order $k = 3, 5, 7$, respectively, and the weight in the objective function is set as $w(\mathbf{x}_t) = \exp\{-\|\mathbf{x}_t\|^2/2\}$ for the weighted penalized estimator and $w(\mathbf{x}_t) = 1$ for the unweighted one.

We note that information criteria, such as the AIC or the BIC, or the cross validation could be adopted to select the sieve order k. However, the results are found similar to fixed k cases and therefore are not reported for space consideration. The results are reported for the SCAD penalty of Fan and Li (2001), with the constant parameter $a = 3.7$ and λ selected by the generalized cross-validation.

It is observed from Tables 2 and 3 ($d = 5$) that the variable selection procedure works reasonably well, even for sample size as small as $n = 200$. When $n = 800$, the selection procedure works close to perfectly. Both the percentage of correct fitting and that of correct nonzeros are almost 100%. Overall, the average numbers of truly nonzero coefficients that are incorrectly estimated to be zero decrease as the sample size increases. Furthermore, the proportions of models correctly fitted increase with the sample size. For DGP 1 and DGP 2 with bounded regression functions, the weighted and unweighted procedures do not seem to differ. However, for DGP 3 and DGP 4, whose regression functions are unbounded, the weighted procedure apparently outperforms the unweighted procedure. For example, in DGP 3, the

Table 3. Variable selection results for DGP 3 and DGP 4, $d = 5$.

		Unweighted					Weighted				
n	k	C	IC	OF	CF	UF	C	IC	OF	CF	UF
						DGP 3					
200	3	79.5	2.0	0.0	56.9	41.1	99.6	0.0	0.1	99.1	0.8
	5	66.5	6.7	0.0	26.2	67.1	98.4	0.0	0.0	96.7	3.3
	7	61.8	10.7	0.0	12.9	76.4	95.2	0.0	0.0	90.3	9.7
800	3	90.5	0.0	0.0	80.9	19.1	100.0	0.0	0.0	100.0	0.0
	5	72.5	0.0	0.0	45.1	54.9	100.0	0.0	0.0	100.0	0.0
	7	66.6	0.0	0.0	33.2	66.8	100.0	0.0	0.0	100.0	0.0
						DGP 4					
200	3	81.6	2.7	0.0	60.5	36.8	96.8	0.9	2.5	91.1	6.4
	5	70.1	3.3	0.0	36.9	59.8	94.3	0.8	2.3	86.4	11.3
	7	67.0	5.4	0.0	28.5	66.1	90.3	0.8	1.6	78.8	19.5
800	3	89.8	0.2	0.0	79.5	20.3	100.0	0.0	0.0	100.0	0.0
	5	79.1	0.1	0.0	58.1	41.8	100.0	0.0	0.0	99.9	0.1
	7	72.9	0.0	0.0	45.7	54.3	100.0	0.0	0.0	99.9	0.1

percentage of correct fitting is 100% for the weighted procedure, while it is only 33.2% for the unweighted one, when $n = 800, k = 7$. This signifies the importance of the weighting when the underlying regression function is potentially unbounded. We further note that the number of sieve terms k does not seem to have much effect on the selection procedure.

We also observe from Tables 4 and 5 that the increase in the dimension of regressors, d, deteriorates the performance of the selection procedure. For example, the percentage of correct fitting for the unweighted selection procedure decreases from 99.2% to 97.0%, when d increases from 5 to 10, for DGP 1 with $n = 800, k = 7$. This is due to the nonparametric nature of our estimation procedure, where the unknown regression function g is approximated using sieve bases. That is to say that this variable selection procedure also suffers from the "curse-of-dimensionality" problem. Therefore, a larger sample size is often required to achieve the same level of accuracy when the dimension of the problem gets larger. This finding highlights the difference between variable selection procedures in semiparametric models and those in parametric models, even though the implementation here resembles a variable selection in parametric models.

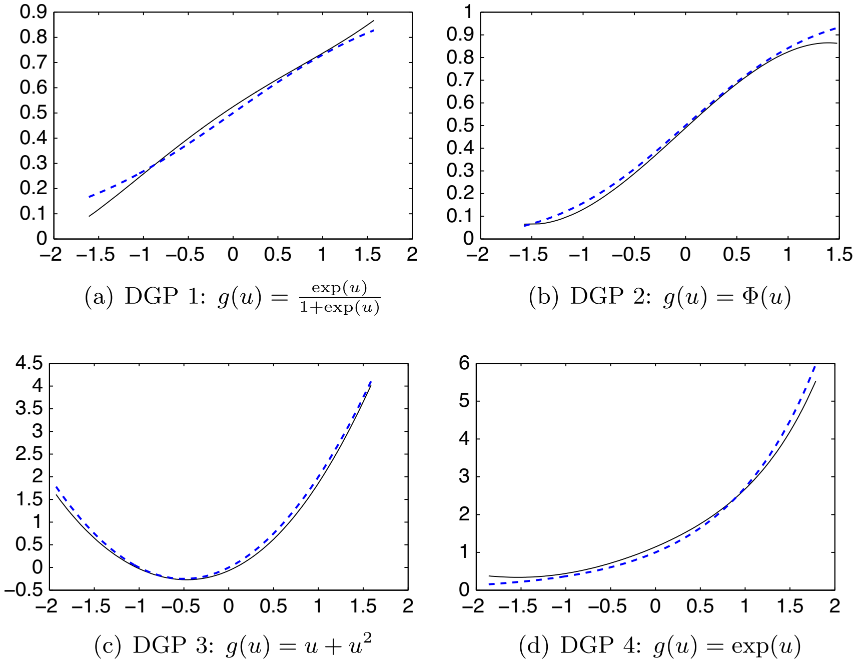
We next evaluate the nonparametric estimator of the function g. We plot in Figure 1 the sieve estimates $\hat{g}(u)$ (dashed line) together with the true function $g(u)$ (solid line), for sample size $n = 200$, $p = 3$, and $k = 5$. It is observed that all four estimated curves are very close to the true curves. This suggests that the sieve estimation of g after variable selection performs satisfactorily for samples with a

TABLE 4. Variable selection results for DGP 1 and DGP 2, $d = 10$.

		Unweighted					Weighted				
n	k	C	IC	OF	CF	UF	C	IC	OF	CF	UF
						DGP 1					
200	3	96.2	11.5	53.4	38.9	7.7	87.4	29.1	68.5	7.5	24.0
	5	92.3	8.9	38.7	45.6	15.3	81.3	22.8	53.2	10.5	35.8
	7	87.2	8.0	30.5	43.1	25.3	70.5	18.1	29.7	14.0	53.7
800	3	100.0	0.5	4.0	96.0	0.0	99.8	11.9	58.8	40.8	0.4
	5	100.0	0.4	3.1	96.9	0.0	99.7	10.8	54.2	45.2	0.6
	7	100.0	0.4	2.9	97.0	0.1	99.6	9.6	50.7	48.4	0.9
						DGP 2					
200	3	96.2	11.2	52.3	40.1	7.6	87.8	28.1	68.3	7.8	23.8
	5	91.4	8.2	38.0	44.8	17.1	81.8	21.7	52.1	13.0	34.6
	7	86.1	7.5	30.8	41.6	27.3	72.9	17.3	31.8	15.8	50.2
800	3	100.0	0.6	4.7	95.3	0.0	99.8	12.2	59.6	39.9	0.5
	5	100.0	0.4	2.7	97.3	0.0	99.8	10.4	52.6	47.0	0.4
	7	100.0	0.2	1.4	98.6	0.0	99.8	9.6	49.6	49.9	0.5

TABLE 5. Variable selection results for DGP 3 and DGP 4, $d = 10$.

		Unweighted					Weighted				
n	k	C	IC	OF	CF	UF	C	IC	OF	CF	UF
						DGP 3					
200	3	71.9	3.5	0.0	40.2	56.3	94.2	0.1	0.8	87.5	11.7
	5	63.8	17.3	0.0	10.4	72.3	86.6	0.0	0.2	73.0	26.8
	7	70.2	38.0	0.0	2.3	59.7	74.9	2.4	0.1	47.3	50.2
800	3	87.5	0.0	0.0	75.0	25.0	100.0	0.0	0.0	100.0	0.0
	5	67.2	0.0	0.0	34.4	65.6	99.9	0.0	0.0	99.8	0.2
	7	59.0	0.0	0.0	18.0	82.0	99.8	0.0	0.0	99.6	0.4
						DGP 4					
200	3	76.5	2.8	0.0	50.2	47.0	86.6	4.0	19.9	53.3	26.8
	5	65.5	8.4	0.0	22.5	69.1	81.0	3.0	13.7	48.3	37.9
	7	64.3	18.3	0.0	10.4	71.3	73.2	5.1	6.1	36.5	53.6
800	3	88.9	0.1	0.0	77.7	22.2	99.2	0.1	0.7	97.6	1.7
	5	74.5	0.0	0.0	49.0	51.0	98.5	0.0	0.3	96.7	3.0
	7	68.6	0.0	0.0	37.2	62.8	98.2	0.0	0.3	96.1	3.6



(a) DGP 1: $g(u) = \frac{\exp(u)}{1+\exp(u)}$

(b) DGP 2: $g(u) = \Phi(u)$

(c) DGP 3: $g(u) = u + u^2$

(d) DGP 4: $g(u) = \exp(u)$

Figure 1. The plots of $\hat{g}(u)$ (dashed line) and the true function $g(u)$ (solid line), for sample size $n = 200$, $p = 3$, and $k = 5$.

typical size. In sum, our proposed approach performs very well in estimating both the parametric and nonparametric components in single index models.

To further evaluate the penalized estimation for the g function whose sieve expansion is sparse, we entertain the following design:

$$\textit{DGP}\quad 5: \qquad y_t = g(\mathbf{x}_t'\boldsymbol{\theta}_0) + e_t,$$

where $g(z) = h_6(z) = H_6(z)/\sqrt{6!}$ and $H_6(\cdot)$ is defined in equation (2.2). Hence, in the orthogonal series expansion of $g(z)$, all coefficients $c_j = 0$ except $c_6 = 1$, that is, the sieve expansion of g presents sparsity. We define the measures C, IC, OF, CF, and UF based on the estimates of $(c_0, \ldots, c_6)^\top$, similar to those defined for $\boldsymbol{\theta}_0$ in the above evaluation. The (weighted) estimation results for $\boldsymbol{\theta}_0$ and g are presented in Table 6, for $d = 5$ and $d_1 = 2$, and $k = 7$. The value of $\boldsymbol{\theta}_0$ is set the same as the preceding designs. Results for other parameter specifications are quite similar and therefore are not presented for space consideration. It is observed from Table 6 that our estimation for both g and $\boldsymbol{\theta}_0$ achieves sparsity with quite good accuracy in the measures calculated. In particular, for sparse function g, with modest sample size we achieve much higher estimation accuracy than that for the parameter $\boldsymbol{\theta}_0$ in an even larger sample size. This finding confirms our theoretical implication that the

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

Table 6. Estimation results for DGP 5, $k = 7$.

	$\boldsymbol{\theta}$					g				
n	C	IC	OF	CF	UF	C	IC	OF	CF	UF
100	96.0	4.1	11.1	84.6	4.3	96.4	0.1	0.7	95.7	3.6
200	98.6	1.6	4.6	93.8	1.6	98.7	0.0	0.2	98.5	1.3
400	99.9	0.3	0.7	99.2	0.1	99.9	0.0	0.0	99.9	0.1

truncation parameter k (sparsity of g) and the dimensionality d of the parameter $\boldsymbol{\theta}_0$ may have different features in our penalized estimation.

5.2. An Empirical Example

This subsection provides a semiparametric analysis of the salaries of major league baseball (MLB) players for the 1987 baseball season. The data set is available in the *R* package *Rfit*. There are $n = 176$ observations in total, with the response variable being the log of the base salary in dollars. There are seven explanatory variables (denoted as $\mathbf{X}$), including log of the number of years experience (*logY*), average wins per year (*aveW*), average losses per year (*aveL*), earned run average (*era*), average games pitched in per year (*aveG*), average number of innings pitched per year (*aveI*), average number of saves per year (*aveS*). To allow for possible nonlinear effects in the determination of the salary, we consider the semiparametric single index model with 119 predictors in total, which include the original seven explanatory variables, their quadratic terms and cubic terms, as well as their interaction terms. All the data are standardized before the analysis.

The proposed screening procedure is first applied to the whole data to detect important variables in the single index model considered in this paper. The sieve order is set as $k = \lfloor n^{1/4} \rfloor = 3$. The BIC criterion selects four variables (denoted as $\mathbf{X}_1$), which are $logY * aveW * aveG$, $logY^2$, $era^2 * aveG$, $logY * aveL * aveG$, as regressors. The estimate of the associated index vector is $(0.927, 0.328, -0.181, -0.030)'$. The penalized estimate of the index parameter through the SCAD penalty is $(0.939, 0.344, 0, 0)'$, which shows that the last two nonlinear predictors are irrelevant. When the EBIC criterion is adopted for variable screening, it leaves us with the first three predictors of $\mathbf{X}_1$ selected by BIC, for that we denote by $\mathbf{X}_2$. The associated penalized estimated index parameter is $(0.9390, 0.3439, 0)'$. Therefore, both the BIC and EBIC selection, married with the SCAD penalization, result in the same single index model for the MLB player salary modeling. Only the first two variables (denoted as $\mathbf{X}_3$) survived from the screening, that is, $logY * aveW * aveG$ and $logY^2$, are relevant in predicting the salary. The estimated link function is plotted in Figure 2.

To evaluate the fitting performance of the proposed model to the data, we consider randomly splitting the whole sample into two subsamples, that is, one subsample with n_1 observations to fit the model and another subsample with $n - n_1$ observations to evaluate the prediction performance. For the purpose of

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

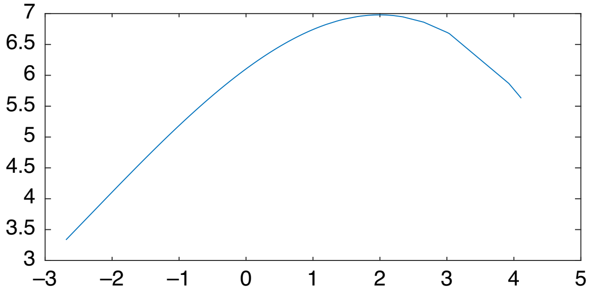


FIGURE 2. The plot of the estimated link function for baseball player salary data with $k = 3$.

TABLE 7. Fitting performance of models M1–M5 to baseball player salary data.

	n_1	M1	M2	M3	M4	M5
	120	0.3389	0.5819	0.1548	0.1915	0.7073
PR^2	140	0.3571	0.6748	0.2552	0.3029	0.7261
	160	0.3520	0.6957	0.3338	0.3519	0.7271
	120	0.4719	0.8734	0.8422	0.8321	0.8007
R^2	140	0.4581	0.8654	0.8334	0.8242	0.7924
	160	0.4630	0.8658	0.8349	0.8261	0.7963

comparison, five models are evaluated. Model M1 is a linear regression of log salary on seven original regressors $\mathbf{X}$. Models M2–M4 are all single index models. Precisely, M2 is constructed with the original seven regressors $\mathbf{X}$, M3 is based on the four selected regressors $\mathbf{X}_1$ after the screening with the BIC, M4 is based on the three selected regressors $\mathbf{X}_2$ after the screening with the EBIC, and M5 is based on the two selected regressors $\mathbf{X}_3$ after the penalized screening. We compute both the in-sample R^2 and the out-of-sample *pseudo* R^2 (PR^2, defined as 1 minus the ratio of out-of-sample mean squared prediction errors over the variance of the log salary), for $n_1 = 120, 140, 160$. The above sample split has been randomly replicated for 500 times and the averaged results over the replications are collected in Table 7. It is observed that the linear model M1 is clearly outperformed by the single index models in both in-sample fits and out-of-sample predictions, revealing the nonlinear feature in the salary determination. Among the single index models, it is observed that the model M2 enjoys the best in-sample fit, followed by M3, M4, and M5. This is consistent with the common knowledge that the more complex the model, the better the in-sample fit. However, when it comes to the out-of-sample prediction, the simplest model M5, that is, the penalized model after the variable screening, has the best prediction accuracy and clearly stands out. This indicates that the penalized estimation after the screening in single index modeling of the baseball player salary data can significantly improve the salary predictions.

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

6. CONCLUSION

This paper considers variable selection and estimation in semiparametric single index models using Hermite polynomial expansion of the unknown link function. The support of the regressor is allowed to be unbounded. The series development offers an approximating linear regression model, based on which shrinkage estimation can be easily implemented with a class of folded concave penalty functions such as the SCAD. The consistency of the proposed selection procedure is established, and asymptotic normality of the index estimator and the link function estimator is proved. The screening procedure is also shown to enjoy the sure screening property. Numerical studies confirm that the proposed procedure enjoys nice finite sample performance.

The current study can be extended in several directions. First, the *i.i.d.* assumption could be relaxed to allow for stationary time series. Second, the regressors are assumed to be exogenous, which can be extended to allow for endogenous regressors. The generalized method of moments, used in Fan and Liao (2014), for example, could be formulated with instrumental variables. Third, the single index structure may be extended to a model with multiple indices. These extensions involve new challenges, and are left for future investigation.

APPENDIX

This appendix contains three parts. Part A collects three auxiliary lemmas, while part B presents the asymptotic results for the coefficients in the linear expansion of the single index model. Part C contains the proof of the main results. Supplementary Material provides the proof for the lemmas and some additional simulation results.

A. Auxiliary Lemmas

Three technical lemmas are shown in this part and their proofs are relegated to the Supplementary Material.

LEMMA A.1. *Suppose that* $\mathbf{u} = (u_1, \ldots, u_d)', \mathbf{v} = (v_1, \ldots, v_d)' \in \mathbb{R}^d$ *and* $\|\mathbf{v}\| = 1$. *Then*

$$H_m(\mathbf{u}'\mathbf{v}) = \sum_{|\mathbf{p}|=m} \binom{m}{\mathbf{p}} \prod_{j=1}^{d} H_{p_j}(u_j) \prod_{j=1}^{d} v_j^{p_j},$$

where $\mathbf{p} = (p_1, \ldots, p_d)$, p_j *for* $j = 1, \ldots, d$ *are all nonnegative integers,* $|\mathbf{p}| = p_1 + \cdots + p_d$ *and* $\binom{m}{\mathbf{p}} = \frac{m!}{\prod_{j=1}^{d} p_j!}$.

LEMMA A.2. *(1) Under Assumption 3.3, for each* $z \in \mathbb{R}$, $g_k(z) \to g(z)$; *meanwhile, for* $\gamma_k(z) = g(z) - g_k(z)$, $\sup_z |\gamma_k(z)|^2 \exp(-z^2/2) = o(k^{-\nu+5/6})$ *and* $\|\gamma_k(z)\|_{L^2}^2 = o(k^{-\nu})$ *as* $k \to \infty$. *(2) Under Assumptions 3.1–3.3,* $\sup_{\boldsymbol{\theta}\in\Theta} \gamma_k^2(\boldsymbol{\theta}'\mathbf{x}_1) = o_P(k^{-\nu})$.

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

LEMMA A.3. *Suppose that for any S with $|S| = s$ and $s^2 = o(n)$ as $n \to \infty$. Under Assumptions 3.1–3.3, we have:*

(1) As $n \to \infty$, $\left\| \frac{1}{n} Z'_{(S)} W Z_{(S)} - \Omega \right\|^2 = O_P(s^2/n)$, where Ω is a square matrix of dimension s having elements $E[\mathcal{H}_{\mathbf{p}}(\mathbf{x}_1)\mathcal{H}_{\mathbf{q}}(\mathbf{x}_1)w(\widetilde{\mathbf{x}}_1)]$, where $\mathbf{p}$ and $\mathbf{q}$ are multiple indices of dimension d varying with $|\mathbf{p}|, |\mathbf{q}| = 0, \ldots, k-1$, whose corresponding j in Definition 2.1 are such that $j \in S$.

(2) As $n \to \infty$, $\left\| \frac{1}{n} Z'_{(S)} W^2 Z_{(S)} - \Psi \right\|^2 = O_P(s^2/n)$, where Ψ is a square matrix of dimension s having elements $E[\mathcal{H}_{\mathbf{p}}(\mathbf{x}_1)\mathcal{H}_{\mathbf{q}}(\mathbf{x}_1)w^2(\widetilde{\mathbf{x}}_1)]$, where $\mathbf{p}$ and $\mathbf{q}$ are the same as that in the above assertion.

(3) As $n \to \infty$, $\left\| \frac{1}{n} Z'_{(S)} W(\mathbf{Y} - Z_{(S)}\boldsymbol{\beta}_0) \right\|^2 = O_P(s/n + sk^{-\nu})$.

(4) Let $Q_{a\times s}$ be a selection matrix for any $a \le s$, that is, each row of Q has one in a place and zeros elsewhere, and all ones are in different columns. Special cases are $Q_{a\times s} = (I_a, 0)$ and the s-vector ℓ_j, where the jth element is 1 and all other elements are zero. Then, $\left\| \frac{1}{n} Q Z'_{(S)} W(\mathbf{Y} - Z_{(S)}\boldsymbol{\beta}_0) \right\|^2 = O_P(a/n + ak^{-\nu})$ for large n. Hence, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty^2 = O_P(1/n + k^{-\nu} + P'_n(\zeta_n))$.

B. Asymptotic Inference for the Linear Coefficients

For any $\boldsymbol{\beta} \in \mathbb{R}^K$, recall the notation $\boldsymbol{\beta}_{S_0}$ given in the first section. Thereby, $\boldsymbol{\beta} = \boldsymbol{\beta}_{S_0} + \boldsymbol{\beta}_{S_0^c}$. In the literature, the subspace $\mathcal{V} = \{\boldsymbol{\beta}_{S_0}, \boldsymbol{\beta} \in \mathbb{R}^K\}$ is called "oracle space" of $\mathbb{R}^K$. Certainly, $\boldsymbol{\beta}_0 \in \mathcal{V}$.

THEOREM B.1. *Let Assumptions 3.1–3.5 hold. Then, there exists a local minimizer $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}'_1, \widehat{\boldsymbol{\beta}}'_2)'$ in optimization (2.8), for which:*

(i) *We have*

$$\lim_{n\to\infty} P(\widehat{\boldsymbol{\beta}}_2 = 0) = 1.$$

(ii) *Let $\widehat{S} = \{j : 1 \le j \le K, \widehat{\beta}_j \ne 0\}$. Then,*

$$\lim_{n\to\infty} P(\widehat{S} = S_0) = 1.$$

(iii) *Suppose that $s_0^2 = o(n)$ and that the $s_0 \times s_0$ matrices Ω and Ψ defined in Lemma A.3 have bounded eigenvalues below from zero and above from infinity uniformly. For any $\boldsymbol{\alpha} \in \mathbb{R}^{s_0}$ with $\|\boldsymbol{\alpha}\| = 1$, we have, as $n \to \infty$, with probability tending to one,*

$$\sqrt{n}(\boldsymbol{\alpha}'\Omega^{-1}\Psi\Omega^{-1}\boldsymbol{\alpha}\sigma_e^2)^{-1/2}\boldsymbol{\alpha}'(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \xrightarrow{d} \mathcal{N}(0,1).$$

This theorem shows that we can obtain a local minimizer from the minimization (2.8) and the minimizer has consistent support and asymptotic normality for the nonzero coefficients. The requirement $s_0^2 = o(n)$ restricts the divergence of the number of nonzero coefficients in the index vector and the orthogonal series expansion.

The uniformly boundedness of eigenvalues for Ω and Ψ is a usual requirement in the literature and is often satisfied in view of their forms. See, for example, Condition A.2 of

Belloni et al. (2015). Lemma A.3 suggests that Ω and Ψ are consistently estimable via their sample counterparts. On the other hand, the residual variance σ_e^2 can be consistently estimated by

$$\widehat{\sigma}_e^2 = \frac{1}{\sum_{t=1}^n w(\widetilde{\mathbf{x}}_t)} \sum_{t=1}^n (y_t - Z_k(\mathbf{x}_t)'\widehat{\boldsymbol{\beta}})^2 w(\widetilde{\mathbf{x}}_t),$$

the proof of which is a routine exercise under the conditions of Theorem B.1. See also Corollary 3.1 in Dong, Linton, and Peng (2021). By virtue of these consistent estimates, the statistical inference can be conducted from the above theorem.

Next, we shall show that under certain additional conditions, the local minimizer in Theorem B.1 is nearly global.

THEOREM B.2. *In addition to Assumptions 3.1–3.5, suppose that* $\sup_{\mathbf{b}:\mathbf{b}_{S_0}\neq 0} \frac{|\mathbf{b}'_{S_0} H_n \mathbf{b}_{S_0^c}|}{\mathbf{b}'_{S_0} H_n \mathbf{b}_{S_0}} \leq C_H < \frac{1}{2}$ *almost surely. Then, the local minimizer in Theorem B.1 satisfies that, for any* $\delta > 0$, *there exists an* $\eta > 0$ *such that*

$$\lim_{n\to\infty} P\left(Q_n(\widehat{\boldsymbol{\beta}}) + \eta < \inf_{\boldsymbol{\beta}\notin\Omega_\delta} Q_n(\boldsymbol{\beta}) \right) = 1, \tag{B.1}$$

where $\Omega_\delta = \{\boldsymbol{\beta} \in \mathbb{R}^K : |\beta_j - \beta_{0j}| \leq \delta, j \leq K\}$ *and* $Q_n(\boldsymbol{\beta})$ *is the objective function in the optimization (2.8).*

Recalling the definition of $\mathbf{b}_S$, the condition on $\sup_{\mathbf{b}:\mathbf{b}_{S_0}\neq 0} \frac{|\mathbf{b}'_{S_0} H_n \mathbf{b}_{S_0^c}|}{\mathbf{b}'_{S_0} H_n \mathbf{b}_{S_0}} \leq C_H < \frac{1}{2}$ requires that the block on the diagonal of H_n corresponding to the support S_0 dominates the block off diagonal corresponding to S_0 and S_0^c, because the numerator $\mathbf{b}'_{S_0} H_n \mathbf{b}_{S_0^c} = \mathbf{b}'_{(S_0)} H_{12} \mathbf{b}_{(S_0^c)}$ while the denominator $\mathbf{b}'_{S_0} H_n \mathbf{b}_{S_0} = \mathbf{b}'_{(S_0)} H_{11} \mathbf{b}_{(S_0)}$, if we partition $H_n = (H_{ij}, i,j = 1,2)$ conformably with the support S_0. One extreme case is that H_{11} is positive definite and $H_{12} = 0$, so that $C_H = 0$. Moreover, given $H_n = Z'Z$, we have $\mathbf{b}'_{S_0} H_n \mathbf{b}_{S_0} = \|Z\mathbf{b}_{S_0}\|^2$, but $\mathbf{b}'_{S_0} H_n \mathbf{b}_{S_0^c} = \langle Z\mathbf{b}_{S_0}, Z\mathbf{b}_{S_0^c}\rangle$ is an inner product. Then, the condition on C_H is shipped to the correlation coefficient of $Z\mathbf{b}_{S_0}$ and $Z\mathbf{b}_{S_0^c}$ and their norms. This condition is in the same spirit as Condition 2 of Lv and Fan (2009).

C. Proofs of the Main Results

Before showing the main results, two lemmas are given and their conditions are illustrated after their proofs. Recall the notation $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_{(S)}$ for index set S with $|S| = s$, and similarly $\boldsymbol{\beta}_{S_0}$ and $\boldsymbol{\beta}_{(S_0)}$ for index set S_0 with $|S_0| = s_0$ defined in Section 3.1.

LEMMA C.1. *Suppose that: (1) There exists a sequence* $a_n = o(\zeta_n)$ *such that* $\|F_{n(S_0)}(\boldsymbol{\beta}_{0(S_0)})\| = O_P(a_n)$. *(2) For any* $\epsilon > 0$, *there exists a constant* $C = C(\epsilon) > 0$ *such that for all large n,* $P(\lambda_{\min}(H_{n(S_0)}(\boldsymbol{\beta}_{0(S_0)})) > C) > 1 - \epsilon$. *Then, there exists a local minimizer* $\widehat{\boldsymbol{\beta}} \in \mathcal{V}$ *of*

$$Q_n(\boldsymbol{\beta}_{S_0}) = L_n(\boldsymbol{\beta}_{S_0}) + \sum_{j\in S_0} P_n(|\beta_j|),$$

such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(a_n + \sqrt{s_0}P_n'(\zeta_n))$*. Moreover, for any arbitrary* $\epsilon > 0$*, the local minimizer* $\widehat{\boldsymbol{\beta}}$ *is strict with probability at least* $1-\epsilon$ *for all large n.*

Proof of Lemma C.1. Define $\rho_n = a_n + \sqrt{s_0}P_n'(\zeta_n)$ and then $\rho_n = o(1)$ by Assumption 3.4. Denote $\mathcal{N}_\tau = \{\boldsymbol{\beta}\in\mathbb{R}^K : \|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_0\| \le \rho_n\tau\}$ for $\tau > 0$. Let $\partial\mathcal{N}_\tau$ be the boundary of $\mathcal{N}_\tau$. Also, define an event

$$A_n(\tau) = \left\{ Q_n(\boldsymbol{\beta}_0) < \min_{\boldsymbol{\beta}\in\partial\mathcal{N}_\tau} Q_n(\boldsymbol{\beta}_{S_0}) \right\}.$$

On the event $A_n(\tau)$, by the continuity of $Q_n(\boldsymbol{\beta})$ with respect to β_j for $j\in S_0$, there exists a local minimizer of $Q_n(\boldsymbol{\beta}_{S_0})$ inside $\mathcal{N}_\tau$. That is, there exists a local minimizer $\widehat{\boldsymbol{\beta}}\in\mathcal{V}$ of $Q_n(\boldsymbol{\beta}_{S_0})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| < \tau\rho_n$. Therefore, it suffices to show that for $\forall\epsilon > 0$, there exists a $\tau > 0$ such that $P(A_n(\tau)) \ge 1-\epsilon$ for all large n.

For any $\boldsymbol{\beta}\in\partial\mathcal{N}_\tau$, viz. $\|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_0\| = \tau\rho_n$, note that

$$\begin{aligned} Q_n(\boldsymbol{\beta}_{S_0}) - Q_n(\boldsymbol{\beta}_0) =& (\boldsymbol{\beta}_{(S_0)} - \boldsymbol{\beta}_{0(S_0)})' F_{n(S_0)}(\boldsymbol{\beta}_{0(S_0)}) + (\boldsymbol{\beta}_{(S_0)} - \boldsymbol{\beta}_{0(S_0)})' H_{n(S_0)}(\boldsymbol{\beta}_{(S_0)} - \boldsymbol{\beta}_{0(S_0)}) \\ &+ \sum_{j\in S_0}[P_n(|\beta_j|) - P_n(|\beta_{0j}|)]. \end{aligned}$$

Invoking the condition $\|F_{n(S_0)}(\boldsymbol{\beta}_{0(S_0)})\| = O_P(a_n)$, for $\forall\epsilon > 0$, there exists a $C_1 > 0$ such that the event A_1 given below satisfies $P(A_1) > 1-\epsilon/2$ for all large n, where

$$A_1 = \{(\boldsymbol{\beta}_{(S_0)} - \boldsymbol{\beta}_{0(S_0)})' F_{n(S_0)}(\boldsymbol{\beta}_{0(S_0)}) \ge -C_1 a_n \|\boldsymbol{\beta}_{(S_0)} - \boldsymbol{\beta}_{0(S_0)}\|\}.$$

Also, by condition (2) and for this ϵ, there exists a C_2 such that $P(A_2) > 1-\epsilon/2$ for all large n, where

$$A_2 = \{(\boldsymbol{\beta}_{(S_0)} - \boldsymbol{\beta}_{0(S_0)})' H_{n(S_0)}(\boldsymbol{\beta}_{(S_0)} - \boldsymbol{\beta}_{0(S_0)}) \le C_2 \|\boldsymbol{\beta}_{(S_0)} - \boldsymbol{\beta}_{0(S_0)}\|^2\}.$$

On the other hand, it follows from Lemma B.1 in Fan and Liao (2014, p. 899) that $\sum_{j\in S_0}[P_n(|\beta_j|) - P_n(|\beta_{0j}|)] \ge -\sqrt{s_0}P_n'(\zeta_n)\|\boldsymbol{\beta}_{(S_0)} - \boldsymbol{\beta}_{0(S_0)}\|$. Hence, for any $\boldsymbol{\beta}\in\partial\mathcal{N}_\tau$, on $A_1\cap A_2$,

$$Q_n(\boldsymbol{\beta}_{S_0}) - Q_n(\boldsymbol{\beta}_0) \ge \rho_n\tau\left(\rho_n\tau C_2 - C_1 a_n - \sqrt{s_0}P_n'(\zeta_n)\right).$$

Since $\rho_n = a_n + \sqrt{s_0}P_n'(\zeta_n)$, we have $C_1 a_n + \sqrt{s_0}P_n'(\zeta_n) \le (C_1+1)\rho_n$. Thus, choosing $\tau > (C_1+1)/C_2$ yields that $Q_n(\boldsymbol{\beta}_{S_0}) - Q_n(\boldsymbol{\beta}_0) > 0$ uniformly on $\boldsymbol{\beta}\in\partial\mathcal{N}_\tau$. It follows that for all large n, with $\tau > (C_1+1)/C_2$, we have $P(A_n(\tau)) > P(A_1\cap A_2) \ge 1-\epsilon$.

We next show that the local minimizer, denoted by $\widehat{\boldsymbol{\beta}}\in\mathcal{V}$, is strict with a probability arbitrarily close to one. For each $h\neq 0$, define

$$\psi(h) = \limsup_{\epsilon\to 0^+} \sup_{(u_1,u_2)\subset O(|h|,\epsilon)} -\frac{P_n'(u_2) - P_n'(u_1)}{u_2 - u_1}.$$

By the concavity, $\psi(\cdot)\ge 0$. Let $\Omega(\boldsymbol{\beta}) = H_{n(S_0)}(\boldsymbol{\beta}_{(S_0)}) - \text{diag}(\psi(\beta_{(S_0),1}),\ldots,\psi(\beta_{(S_0),s_0}))$, for any $\boldsymbol{\beta}\in\mathcal{N}_\tau$, where we denote $\boldsymbol{\beta}_{(S_0)} = (\beta_{(S_0),1},\ldots,\beta_{(S_0),s_0})'$. It suffices to show that $\Omega(\widehat{\boldsymbol{\beta}})$ is positive definite with probability arbitrarily close to unity.

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

On the event $A_3 = \{\phi(\widehat{\boldsymbol{\beta}}_{(S_0)}) \leq \sup_{\boldsymbol{\beta}_{(S_0)} \in O(\boldsymbol{\beta}_{0(S_0)}, c\zeta_n)} \phi(\boldsymbol{\beta}_{(S_0)})\}$, where $\widehat{\boldsymbol{\beta}}_{(S_0)}$ is the counterpart of $\boldsymbol{\beta}_{(S_0)}$, and c is the same in (iv) of Assumption 3.4, we have

$$\max_{j \leq s_0} \psi(\widehat{\beta}_{(S_0),j}) \leq \phi(\widehat{\boldsymbol{\beta}}_{(S_0)}) \leq \sup_{\boldsymbol{\beta}_{(S_0)} \in O(\boldsymbol{\beta}_{0(S_0)}, c\zeta_n)} \phi(\boldsymbol{\beta}_{(S_0)}).$$

Let $A_4 = \{\lambda_{\min}(H_{n(S_0)}(\boldsymbol{\beta}_{0(S_0)})) > C_2\}$. Then, for any $\mathbf{u} \in \mathbb{R}^{s_0}$ with $\|\mathbf{u}\| = 1$, it follows from (iv) of Assumption 3.4 that

$$\begin{aligned}\mathbf{u}'\Omega(\widehat{\boldsymbol{\beta}})\mathbf{u} =& \mathbf{u}'H_{n(S_0)}(\widehat{\boldsymbol{\beta}}_{(S_0)})\mathbf{u} - \mathbf{u}'\mathrm{diag}(\psi(\widehat{\beta}_{(S_0),1}), \ldots, \psi(\widehat{\beta}_{(S_0),s_0}))\mathbf{u} \\ \geq& C_2 - \sup_{\boldsymbol{\beta}_{(S_0)} \in O(\boldsymbol{\beta}_{0(S_0)}, c\zeta_n)} \phi(\boldsymbol{\beta}_{(S_0)}) \geq C_2/2,\end{aligned}$$

on the event $A_3 \cap A_4$ for all large n.

Finally, we are about to show that $P(A_3 \cap A_4) \geq 1 - \epsilon$. Indeed, due to $\rho_n = o(\zeta_n)$, $P(A_3) \geq P(\widehat{\boldsymbol{\beta}}_{(S_0)} \in O(\boldsymbol{\beta}_{0(S_0)}, c\zeta_n)) \geq 1 - \epsilon/2$ for all large n. Also, $P(A_4) \geq 1 - \epsilon/2$ due to the condition (2), the assertion then follows. □

The oracle consistency in Lemma C.1 is derived based on the knowledge of S_0, the support of $\boldsymbol{\beta}_0$; the lemma has independent interest because the condition (2) weakens Assumption 3.5(iv). To make the result useful, it is desirable to show that the local minimizer of Q_n restricted on $\mathcal{V}$ is also a minimizer of Q_n on $\mathbb{R}^K$.

LEMMA C.2. *Additional to the conditions in Lemma C.1, suppose that with probability approaching one, for $\widehat{\boldsymbol{\beta}} \in \mathcal{V}$ in Lemma C.1, there exists a neighborhood $O_1 \subset \mathbb{R}^K$ of $\widehat{\boldsymbol{\beta}}$ such that for all $\boldsymbol{\beta} \in O_1$ but $\boldsymbol{\beta} \notin \mathcal{V}$, we have*

$$L_n(\boldsymbol{\beta}_{S_0}) - L_n(\boldsymbol{\beta}) < \sum_{j \notin S_0} P_n(|\beta_j|). \tag{C.1}$$

Then, (i) With probability close to unity arbitrarily, the $\widehat{\boldsymbol{\beta}} \in \mathcal{V}$ is a local minimizer in $\mathbb{R}^K$ of $Q_n(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta}) + \sum_{j=1}^K P_n(|\beta_j|)$. (ii) For $\forall \epsilon > 0$, the local minimizer $\widehat{\boldsymbol{\beta}} \in \mathcal{V}$ is strict with probability at least $1 - \epsilon$ for all large n.

Proof of Lemma C.2. Recall that $\widehat{\boldsymbol{\beta}} \in \mathcal{V}$ is a local minimizer of $Q_n(\boldsymbol{\beta}_{S_0})$. Hence, there is a small neighborhood O_1 of $\widehat{\boldsymbol{\beta}}$, such that for any $\boldsymbol{\beta} \in O_1$ with $\boldsymbol{\beta} \notin \mathcal{V}$, we have $Q_n(\widehat{\boldsymbol{\beta}}) \leq Q_n(\boldsymbol{\beta}_{S_0})$. However, by the condition of (C.1),

$$Q_n(\boldsymbol{\beta}_{S_0}) - Q_n(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta}_{S_0}) - L_n(\boldsymbol{\beta}) - \sum_{j \notin S_0} P_n(|\beta_j|) < 0. \tag{C.2}$$

This means $Q_n(\widehat{\boldsymbol{\beta}}) < Q_n(\boldsymbol{\beta})$, yielding the first assertion, while, from which and the last statement of Lemma C.1, the second assertion is also implied. □

Proof of Theorem B.1. (i) As shown in Lemma C.1, if $Q_n(\boldsymbol{\beta})$ has a local minimizer $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2')'$, then $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$ with probability arbitrarily close to one for large n, which implies the assertion (i) and $P(\widehat{S} \subset S_0) \to 1$.

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

On the other hand,

$$\begin{aligned}P(S_0 \not\subset \widehat{S}) =& P(\exists j \in S_0, \widehat{\beta}_j = 0) \leq P(\exists j \in S_0, |\beta_{0j} - \widehat{\beta}_j| \geq |\beta_{0j}|)\\ \leq& P(\max_j |\beta_{0j} - \widehat{\beta}_j| \geq \zeta_n) \leq P(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \geq \zeta_n) = o(1),\end{aligned}$$

implying $P(S_0 \subset \widehat{S}) \to 1$. Accordingly, $P(S_0 = \widehat{S}) \to 1$ which proves (ii).

(iii) Let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2')'$ be the local minimizer of $Q_n(\boldsymbol{\beta})$ where $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$ with probability arbitrarily close to one. Define $P_n'(|\widehat{\boldsymbol{\beta}}_1|) := (P_n'(|\widehat{\beta}_{11}|), \ldots, P_n'(|\widehat{\beta}_{1s_0}|))'$ and $\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_1) := (\mathrm{sgn}(\widehat{\beta}_{11}), \ldots, \mathrm{sgn}(\widehat{\beta}_{1s_0}))'$.

By the Karush–Kuhn–Tucker (KKT) condition,

$$F_{n(S_0)}(\widehat{\boldsymbol{\beta}}_1) = -P_n'(|\widehat{\boldsymbol{\beta}}_1|) \diamond \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_1),$$

where the operator $\diamond$ is the product in elementwise. Observe that

$$F_{n(S_0)}(\widehat{\boldsymbol{\beta}}_1) = F_{n(S_0)}(\boldsymbol{\beta}_1) + H_{n(S_0)}(\boldsymbol{\beta}_1)(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1),$$

which further implies

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 &= H_{n(S_0)}(\boldsymbol{\beta}_1)^{-1}[F_{n(S_0)}(\widehat{\boldsymbol{\beta}}_1) - F_{n(S_0)}(\boldsymbol{\beta}_1)]\\ &= -H_{n(S_0)}(\boldsymbol{\beta}_1)^{-1}[F_{n(S_0)}(\boldsymbol{\beta}_1) + P_n'(|\widehat{\boldsymbol{\beta}}_1|) \diamond \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_1)]\\ &= \frac{2}{n} H_{n(S_0)}(\boldsymbol{\beta}_1)^{-1} Z_{(S_0)}' W(\mathbf{e} + \boldsymbol{\gamma}) - H_{n(S_0)}(\boldsymbol{\beta}_1)^{-1} P_n'(|\widehat{\boldsymbol{\beta}}_1|) \diamond \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_1)\\ &= \frac{1}{n}\left(\frac{1}{n} Z_{(S_0)}' W Z_{(S_0)}\right)^{-1} Z_{(S_0)}' W\mathbf{e} + \frac{1}{n}\left(\frac{1}{n} Z_{(S_0)}' W Z_{(S_0)}\right)^{-1} Z_{(S_0)}' W\boldsymbol{\gamma}\\ &\quad - \left(\frac{1}{n} Z_{(S_0)}' W Z_{(S_0)}\right)^{-1} P_n'(|\widehat{\boldsymbol{\beta}}_1|) \diamond \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_1)\\ &= \frac{1}{n}\Omega^{-1} Z_{(S_0)}' W\mathbf{e}(1 + o_P(1)) + \frac{1}{n}\left(\frac{1}{n} Z_{(S_0)}' W Z_{(S_0)}\right)^{-1} Z_{(S_0)}' W\boldsymbol{\gamma}\\ &\quad - \frac{1}{n}\left(\frac{1}{n} Z_{(S_0)}' W Z_{(S_0)}\right)^{-1} P_n'(|\widehat{\boldsymbol{\beta}}_1|) \diamond \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_1),\end{aligned}$$

by the definitions of $H_{n(S_0)}$ and $F_{n(S_0)}$ and Lemma A.3. The normality shall be derived from the first term.

Denote by $Z_{k(S_0)}(\mathbf{x}_t)$, the vector $Z_k(\mathbf{x}_t)$ eliminating all elements whose subscripts are not in S_0, so that $Z_{k(S_0)}(\mathbf{x}_t)$ is a s_0-vector. That is, $Z_{k(S_0)}(\mathbf{x}_t)$ are all the columns of Z_{S_0}'. Hence,

$$\frac{1}{n}\alpha'\Omega^{-1} Z_{S_0}' W\mathbf{e} = \frac{1}{n}\alpha'\Omega^{-1}\sum_{t=1}^{n} Z_{k(S_0)}(\mathbf{x}_t) w(\widetilde{\mathbf{x}}_t) e_t,$$

which has variance

$$\alpha'\Omega^{-1}\left(\frac{1}{n} Z_{(S_0)}' W^2 Z_{(S_0)}\right)\Omega^{-1}\alpha\sigma_e^2 = \alpha\Omega^{-1}\Psi\Omega^{-1}\alpha\sigma_e^2(1 + o_P(1)),$$

by Lemma A.3 again. Therefore, it follows from the conditional Lindeberg central limit theorem that

$$\sqrt{n}(\alpha'\Omega^{-1}\Psi\Omega^{-1}\alpha\sigma_e^2)^{-1/2}\alpha'\Omega^{-1}Z'_{(S_0)}W\mathbf{e} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$.

It remains to show that $n^{-1/2}Z'_{(S_0)}W\boldsymbol{\gamma} = o_P(1)$ and $n^{1/2}P_n'(|\widehat{\boldsymbol{\beta}}_1|)\diamond \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_1) = o_P(1)$.

Note by the boundedness of the elements in $Z'_{(S_0)}W$, $n^{-1/2}\|Z'_{(S_0)}W\boldsymbol{\gamma}\| \le n^{-1/2}\|Z'_{(S_0)}W\|$ $\|\boldsymbol{\gamma}\| = O_P(\sqrt{s_0 n}\|\gamma_k(z)\|) = o_P(1)$ due to Assumption 3.5.

Similar to Lemma C.2 of Fan and Liao (2014), we may show that

$$\|P_n'(|\widehat{\boldsymbol{\beta}}_1|)\diamond \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_1)\| = O_P\left(\sup_{\|\boldsymbol{\beta}_{S_0}-\boldsymbol{\beta}_1\|\le\zeta_n/4}\phi(\boldsymbol{\beta}_{S_0})\sqrt{s_0\log(K)/n}+\sqrt{s_0}\|\gamma_k(u)\|+\sqrt{s_0}P_n'(\zeta_n)\right).$$

Thus, $n^{1/2}P_n'(|\widehat{\boldsymbol{\beta}}_1|)\diamond \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_1) = o_P(1)$ due to Assumption 3.5 again, which finishes the proof. □

Proof of Theorem B.2. First, using the mean value theorem,

$$\begin{aligned}\sum_{j\in S_0}P_n(|\widehat{\beta}_j|) &\le \sum_{j\in S_0}P_n(|\beta_{0j}|)+\sum_{j\in S_0}P_n'(|\beta_{0j}^*|)|\widehat{\beta}_j-\beta_{0j}|\\ &\le s_0\max_{j\in S_0}P_n(|\beta_{0j}|)+\sum_{j\in S_0}P_n'(\zeta_n)|\widehat{\beta}_j-\beta_{0j}|\\ &\le s_0\max_{j\in S_0}P_n(|\beta_{0j}|)+\sqrt{s_0}P_n'(\zeta_n)\|\widehat{\boldsymbol{\beta}}_{(S_0)}-\boldsymbol{\beta}_1\|.\end{aligned}$$

Second,

$$\begin{aligned}Q_n(\boldsymbol{\beta}) &= L_n(\boldsymbol{\beta}_0)+(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-2(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'F_n(\boldsymbol{\beta}_0)+\sum_{j=1}^{K}P_n(|\beta_j|),\\ Q_n(\widehat{\boldsymbol{\beta}}) &= L_n(\boldsymbol{\beta}_0)+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'H_n(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)-2(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'F_n(\boldsymbol{\beta}_0)+\sum_{j\in S_0}P_n(|\widehat{\beta}_j|),\end{aligned}$$

which gives

$$\begin{aligned}Q_n(\boldsymbol{\beta})-Q_n(\widehat{\boldsymbol{\beta}}) = {}&(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'H_n(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\\ &-2(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'F_n(\boldsymbol{\beta}_0)+2(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'F_n(\boldsymbol{\beta}_0)\\ &+\sum_{j=1}^{K}P_n(|\beta_j|)-\sum_{j\in S_0}P_n(|\widehat{\beta}_j|).\end{aligned}$$

For the first term, using the identity $\mathbf{u} = \mathbf{u}_{S_0}+\mathbf{u}_{S_0^c}$,

$$\begin{aligned}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0) = {}&(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'_{S_0}H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{S_0}+(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'_{S_0^c}H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{S_0^c}\\ &+2(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'_{S_0}H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{S_0^c}\\ \ge{}&(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'_{S_0}H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{S_0}+2(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'_{S_0}H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{S_0^c}\end{aligned}$$

$$
\begin{aligned}
&\geq (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'_{S_0}H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{S_0}\left[1-2\frac{|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'_{S_0}H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{S_0^c}|}{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'_{S_0}H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{S_0}}\right]\\
&\geq (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'_{S_0}H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{S_0}\left[1-2\sup_{\mathbf{u}\notin\Omega_\delta}\frac{|\mathbf{u}'_{S_0}H_n\mathbf{u}_{S_0^c}|}{\mathbf{u}'_{S_0}H_n\mathbf{u}_{S_0}}\right]\\
&\geq (1-2C_H)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'_{S_0}H_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{S_0}\\
&= (1-2C_H)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'_{(S_0)}H_{n(S_0)}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{(S_0)}\\
&\geq (1-2C_H)\lambda_{\min}(H_{n(S_0)})\|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{(S_0)}\|^2\\
&\geq (1-2C_H)C_1\|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{(S_0)}\|^2,
\end{aligned}
$$

where $(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{(S_0)}$ is a short version of $(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{S_0}$ deleting all zeros at $j\notin S_0$. Therefore,

$$
\begin{aligned}
Q_n(\boldsymbol{\beta})-Q_n(\widehat{\boldsymbol{\beta}}) \geq{}& (1-2C_H)C_1\|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{(S_0)}\|^2-(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'H_n(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\\
&-2(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'F_n(\boldsymbol{\beta}_0)+2(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'F_n(\boldsymbol{\beta}_0)-\sum_{j\in S_0}P_n(|\widehat{\beta}_j|)\\
>{}& (1-2C_H)C_1\|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{(S_0)}\|^2-\lambda_{\max}(H_n)\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2\\
&-2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\|F_n(\boldsymbol{\beta}_0)\|-2\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|\|F_n(\boldsymbol{\beta}_0)\|\\
&-s_0\max_{j\in S_0}P_n(|\beta_{0j}|)-\sqrt{s_0}P_n'(\zeta_n)\|\widehat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}_1\|\\
:={}& (1-2C_H)C_1\|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)_{(S_0)}\|^2-\xi_n,
\end{aligned}
$$

where $0\leq\xi_n=o_P(1)$ by noting that $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=o_P(1)$, $\|F_n(\boldsymbol{\beta}_0)\|=o_P(1)$, and by the condition of $s_0\max_{j\in S_0}P_n(|\beta_{0j}|)=o(1)$ which is implied from Assumption 3.5. Then, for any $\delta>0$, we take $\eta=(1-2C_H)C_1\delta^2/2>0$, so that

$$
\begin{aligned}
&P\left(\inf_{\boldsymbol{\beta}\notin\Omega_\delta}Q_n(\boldsymbol{\beta})-Q_n(\widehat{\boldsymbol{\beta}})>\eta\right)\\
\geq{}&P\left(\inf_{\boldsymbol{\beta}\notin\Omega_\delta}\lambda_{\min}(H_n)\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2\geq\eta+\xi_n\right)\\
\geq{}&P\left(\xi_n\leq(1-2C_H)C_1\delta^2/2\right)\to 1,
\end{aligned}
$$

as $n\to\infty$. □

Proof of Theorem 3.1. (*i*) This is easily obtained from Theorem B.1. (*ii*) Observe that, given $c_i\neq 0$ and $\widehat{c}_i\neq 0$,

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1 &= \frac{1}{\widehat{c}_i\hat{\theta}_{01}^{i-1}}A_iR_i\widehat{\boldsymbol{\beta}}_1-\frac{1}{c_i\theta_{01}^{i-1}}A_iR_i\boldsymbol{\beta}_1\\
&= \frac{1}{\widehat{c}_i\hat{\theta}_{01}^{i-1}}\left(A_iR_i\widehat{\boldsymbol{\beta}}_1-\frac{\widehat{c}_i\hat{\theta}_{01}^{i-1}}{c_i\theta_{01}^{i-1}}A_iR_i\boldsymbol{\beta}_1\right)\\
&= \frac{1}{\widehat{c}_i\hat{\theta}_{01}^{i-1}}\left(A_iR_i(\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)-\frac{c_i\theta_{01}^{i-1}-\widehat{c}_i\hat{\theta}_{01}^{i-1}}{c_i\theta_{01}^{i-1}}A_iR_i\boldsymbol{\beta}_1\right),
\end{aligned}
$$

or equivalently,

$$\begin{aligned}\widehat{c}_i\hat{\theta}_{01}^{i-1}(\widehat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1) &= A_iR_i(\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)-(c_i\theta_{01}^{i-1}-\widehat{c}_i\hat{\theta}_{01}^{i-1})\boldsymbol{\theta}_1\\ &= A_iR_i(\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)-\left(\frac{\beta_{i_1}}{\theta_{01}}-\frac{\widehat{\beta}_{i_1}}{\widehat{\theta}_{01}}\right)\boldsymbol{\theta}_1\\ &= A_iR_i(\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)+\frac{\widehat{\beta}_{i_1}-\beta_{i_1}}{\theta_{01}}\boldsymbol{\theta}_1-\frac{\widehat{\beta}_{i_1}(\widehat{\theta}_{01}-\theta_{01})}{\theta_{01}\widehat{\theta}_{01}}\boldsymbol{\theta}_1\\ &= \left[A_iR_i+\frac{1}{\theta_{01}}\boldsymbol{\theta}_1\ell_{i_1}'\right](\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)-\frac{\widehat{\beta}_{i_1}}{\theta_{01}\widehat{\theta}_{01}}\boldsymbol{\theta}_1\ell_1'(\widehat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1),\end{aligned}$$

where ℓ_{i_1} is a s_0-vector whose i_1th element is 1 and elsewhere zero, while ℓ_1 is a $|J|$-vector whose first element is 1 and elsewhere zero. We further write

$$\left[\widehat{c}_i\hat{\theta}_{01}^{i-1}I_{|J|}+\frac{\widehat{\beta}_{i_1}}{\theta_{01}\widehat{\theta}_{01}}\boldsymbol{\theta}_1\ell_1'\right](\widehat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1)=\left[A_iR_i+\frac{1}{\theta_{01}}\boldsymbol{\theta}_1\ell_{i_1}'\right](\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1).$$

Here, $\widehat{c}_i\hat{\theta}_{01}^{i-1}=\widehat{\beta}_{i_1}/\hat{\theta}_{01}$ and the matrix $I_{|J|}+\theta_{01}^{-1}\boldsymbol{\theta}_1\ell_1'$ is lower triangular with diagonal elements $(2,1,\ldots,1)$, and hence it is invertible and the inverse can be easily obtained. We then have

$$\begin{aligned}\frac{\widehat{\beta}_{i_1}}{\widehat{\theta}_{01}}\alpha'(\widehat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1) &= \alpha'\left[I_{|J|}+\frac{1}{\theta_{01}}\boldsymbol{\theta}_1\ell_1'\right]^{-1}\left[A_iR_i+\frac{1}{\theta_{01}}\boldsymbol{\theta}_1\ell_{i_1}'\right](\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)\\ &:= \alpha'B_{ni}(\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1), \qquad \textbf{(C.3)}\end{aligned}$$

where $B_{ni}:=\left[I_{|J|}+\frac{1}{\theta_{01}}\boldsymbol{\theta}_1\ell_1'\right]^{-1}\left[A_iR_i+\frac{1}{\theta_{01}}\boldsymbol{\theta}_1\ell_{i_1}'\right]$.

It then follows from the proof of Theorem B.1 that

$$\begin{aligned}\frac{\widehat{\beta}_{i_1}}{\widehat{\theta}_{01}}\alpha'(\widehat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1) &= \alpha'B_{ni}(\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)\\ &= \frac{1}{n}\alpha'B_{ni}\Omega^{-1}Z_{(S_0)}'W\mathbf{e}(1+o_P(1))\\ &\quad+\alpha'B_{ni}\frac{1}{n}\left(\frac{1}{n}Z_{(S_0)}'WZ_{(S_0)}\right)^{-1}Z_{(S_0)}'W\boldsymbol{\gamma}\\ &\quad-\alpha'B_{ni}\left(\frac{1}{n}Z_{(S_0)}'WZ_{(S_0)}\right)^{-1}P_n'(|\widehat{\boldsymbol{\beta}}_1|)\diamond\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_1),\end{aligned}$$

where Ω is given in Lemma A.3.

Here, $\sqrt{n}\frac{\widehat{\beta}_{i_1}}{\widehat{\theta}_{01}}\alpha'(\widehat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1)$ has leading term $\frac{1}{\sqrt{n}}\alpha'B_{ni}\Omega^{-1}Z_{(S_0)}'W\mathbf{e}$, for which, similar to the proof of Theorem B.1, we have

$$\sigma_{ni}^{-1}\frac{1}{\sqrt{n}}\alpha'B_{ni}\Omega^{-1}Z_{(S_0)}'W\mathbf{e}\xrightarrow{d}\mathcal{N}(0,1)$$

as $n\to\infty$, where $\sigma_{ni}^2=\alpha'B_{ni}\Omega^{-1}\Psi\Omega^{-1}B_{ni}'\alpha\sigma_e^2$ and Ψ is also given in Lemma A.3.

All the other terms are negligible due to the same reason in the proof of Theorem B.1. The proof is then finished. □

Proof of Theorem 3.2. (*i*) This is the implication of Theorem B.1. (*ii*) Firstly, we show $n^{-1}\|\widehat{U}'\widehat{W}\widehat{U}-U'W_0U\| = O_P(d_1^{3/2}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|) = o_P(1)$. In fact, the matrix $n^{-1}(\widehat{U}'\widehat{W}\widehat{U}-U'W_0U)$ has elements $n^{-1}\sum_{t=1}^n[h_i(\widehat{\boldsymbol{\theta}}'\mathbf{x}_t)h_j(\widehat{\boldsymbol{\theta}}'\mathbf{x}_t)w(\widehat{\boldsymbol{\theta}}'\bar{\mathbf{x}}_t)-h_i(\boldsymbol{\theta}_0'\mathbf{x}_t)h_j(\boldsymbol{\theta}_0'\mathbf{x}_t)w(\boldsymbol{\theta}_0'\bar{\mathbf{x}}_t)]$ for $i,j\in\widehat{I}$. Also, $h_i'(z)=\sqrt{i}h_{i-1}(z)$ and $\sup_{i,j,z}|h_i(z)h_j(z)w(z)|<\infty$; remember $w(z)=1$ if z has bounded support. The assertion follows immediately by the mean value theorem. In addition, by the same approach, we have $n^{-1}\|\widehat{U}'\widehat{W}-U'W_0\| = O_P(d_1\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|) = o_P(1)$ as $n\to\infty$.

Let Ξ be the matrix with elements $E[h_i(\boldsymbol{\theta}_0'\mathbf{x}_t)h_j(\boldsymbol{\theta}_0'\mathbf{x}_t)w(\boldsymbol{\theta}_0'\bar{\mathbf{x}}_t)]$ for $i,j\in\widehat{I}$. Then, by the Law of Large Numbers, we have $\|n^{-1}U'W_0U-\Xi\|^2 = O_P(d_1^2/n) = o_P(1)$.

Observe that

$$\begin{aligned}\widetilde{g}(z)-g(z) =&\Phi_{(\widehat{I})}(z)'(\widetilde{c}_{\widehat{I}}-c_{\widehat{I}})-\gamma_k(z)\\ =&\frac{1}{n}\Phi_{(\widehat{I})}(z)'\Xi^{-1}U'W_0(\mathbf{e}+\boldsymbol{\gamma})(1+o_P(1))-\gamma_k(z).\end{aligned}$$

The leading term of $\frac{\sqrt{n}}{\|\Phi_{(\widehat{I})}(z)\|}[\widetilde{g}(z)-g(z)]$ is $\frac{1}{\sqrt{n}}\overline{\Phi}_{(\widehat{I})}(z)'\Xi^{-1}U'W_0\mathbf{e}$ from which we shall derive the normality, where we define $\overline{\Phi}_{(\widehat{I})}(z)=\Phi_{(\widehat{I})}(z)/\|\Phi_{(\widehat{I})}(z)\|$ a unit vector. Note that its conditional variance $\frac{1}{n}\overline{\Phi}_{(\widehat{I})}(z)'\Xi^{-1}U'W_0^2U\Xi^{-1}\overline{\Phi}_{(\widehat{I})}(z)\sigma_e^2=\overline{\Phi}_{(\widehat{I})}(z)'\Xi^{-1}\Sigma\Xi^{-1}\overline{\Phi}_{(\widehat{I})}(z)$ $\sigma_e^2(1+o_P(1))$, where we define Σ to be the asymptotic matrix of $\frac{1}{n}U'W_0^2U$, which has elements $E[h_i(\boldsymbol{\theta}_0'\mathbf{x}_t)h_j(\boldsymbol{\theta}_0'\mathbf{x}_t)w^2(\boldsymbol{\theta}_0'\bar{\mathbf{x}}_t)]$ for $i,j\in\widehat{I}$. By Lemma A.3, $\|n^{-1}U'W_0^2U-\Sigma\|^2 = O_P(d_1^2/n)=o_P(1)$. Let $\sigma_{ni}^2=\overline{\Phi}_{(\widehat{I})}(z)'\Xi^{-1}\Sigma\Xi^{-1}\overline{\Phi}_{(\widehat{I})}(z)\sigma_e^2$. Then, by Assumption 3.1 and the standard central limit theorem, $\sigma_n^{-1}\frac{1}{\sqrt{n}}\overline{\Phi}_{(\widehat{I})}(z)'\Xi^{-1}U'W_0\mathbf{e}\xrightarrow{d}\mathcal{N}(0,1)$.

Meanwhile, by Assumption 3.5, $\frac{1}{\sqrt{n}}\overline{\Phi}_{(\widehat{I})}(z)'\Xi^{-1}U'W_0\boldsymbol{\gamma}=o_P(1)$. In fact, by the uniform boundedness of $h_i(z)w(z)$ and the eigenvalues of Ξ, we have $\frac{1}{\sqrt{n}}|\overline{\Phi}_{(\widehat{I})}(z)'\Xi^{-1}U'W_0\boldsymbol{\gamma}|\le \frac{1}{\sqrt{n}}\|\boldsymbol{\gamma}\|=o_P(\sqrt{n}\|\gamma_k(z)\|)=o_P(1)$. The assertion holds in view of the condition on $\gamma_k(z)$ because $\|\Phi_{(\widehat{I})}(z)\|^2=O_P(|\widehat{I}|)=O_P(d_1)$. □

Proof of Theorem 4.1 (Lower Bound). Denote $H_B=I_n-Z_B(Z_B'Z_B)^{-1}Z_B'$ for any $B\subset\{1,\ldots,d\}$ and Z_B is given by (4.6). Then, from equation (4.6), $n\widehat{\sigma}_B^2=\|\widehat{e}_B\|^2=\|H_B\mathbf{Y}\|^2$.

On the other hand, premultiplying H_B on equation (4.10) gives $H_B\mathbf{Y}=H_B\tilde{Z}_{B(\ell)}\boldsymbol{\beta}_\ell+H_B\gamma_{B(\ell)}+H_Be_{B(\ell)}$, from which the OLS scheme gives

$$\widehat{\boldsymbol{\beta}_\ell}=(\tilde{Z}_{B(\ell)}'H_B\tilde{Z}_{B(\ell)})^{-1}\tilde{Z}_{B(\ell)}'H_B\mathbf{Y}.$$

Hence, $\widehat{e}_{B(\ell)}=H_{B(\ell)}H_BY$ where $H_{B(\ell)}=I_n-P_{\ell B}$ with $P_{\ell B}=H_B\tilde{Z}_{B(\ell)}(\tilde{Z}_{B(\ell)}'H_B\tilde{Z}_{B(\ell)})^{-1}$ $\tilde{Z}_{B(\ell)}'H_B$. Then we have

$$\begin{aligned}n\widehat{\sigma}_B^2-n\widehat{\sigma}_{B(\ell)}^2 =&\|\widehat{e}_B\|^2-\|\widehat{e}_{B(\ell)}\|^2=\|H_B\mathbf{Y}\|^2-\|H_{B(\ell)}H_B\mathbf{Y}\|^2\\ =&\|P_{\ell B}\mathbf{Y}\|^2=\widehat{\boldsymbol{\beta}_\ell}'(\tilde{Z}_{B(\ell)}'H_B\tilde{Z}_{B(\ell)})\widehat{\boldsymbol{\beta}_\ell}.\end{aligned}$$

It follows that $\max_{\ell\in B^c}(n\widehat{\sigma}_B^2-n\widehat{\sigma}_{B(\ell)}^2)=\max_{\ell\in B^c}\|P_{\ell B}\mathbf{Y}\|^2\ge\max_{\ell\in J\backslash B}\|P_{\ell B}\mathbf{Y}\|^2$, where J is the index set of all nonzero components $\boldsymbol{\theta}_1$.

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

In the oracle model $y_t = g(\mathbf{x}_t'\boldsymbol{\theta}_0) + \epsilon_t = g(\mathbf{x}_{t,J}'\boldsymbol{\theta}_1) + \epsilon_t$, where $\|\boldsymbol{\theta}_1\| = 1$, $t = 1, \ldots, n$, similar to (4.6), we may write

$$y_t = Z_k(\mathbf{x}_{t,J})'\boldsymbol{\beta}_J + \gamma_k(\mathbf{x}_{t,J}'\boldsymbol{\theta}_1) + \epsilon_t, \tag{C.4}$$

where $\gamma_k(\cdot)$ is given in (2.4). Write these equations in matrix form $\mathbf{Y} = Z_J\boldsymbol{\beta}_J + \boldsymbol{\gamma}_{k,J} + \boldsymbol{\epsilon}$. It follows that

$$\begin{aligned}\|P_{\ell B}\mathbf{Y}\| =& \|P_{\ell B}(Z_J\boldsymbol{\beta}_J + \boldsymbol{\gamma}_{k,J} + \boldsymbol{\epsilon})\| \\ \geq& \|P_{\ell B}Z_J\boldsymbol{\beta}_J\| - \|P_{\ell B}\boldsymbol{\gamma}_{k,J}\| - \|P_{\ell B}\boldsymbol{\epsilon}\|.\end{aligned}$$

As $P_{\ell B}$ is an orthogonal project matrix, $\|P_{\ell B}\boldsymbol{\gamma}_{k,J}\| \leq \|\boldsymbol{\gamma}_{k,J}\| = O_P(\sqrt{n}k^{-s/2})$. Moreover, by Proposition 3 of Zhang (Zhang, 2010) with $x = C(\log n)^{\tau}$ for some $0 < \tau < 1$, we have

$$P\left(\frac{\boldsymbol{\epsilon}'P_{\ell B}\boldsymbol{\epsilon}}{K_B C_{\epsilon}} \geq \frac{1+x}{(1-2/(e^{x/2}\sqrt{1+x}-1))_+^2}\right) \leq \exp(-K_B x/2)\,(1+x)^{K_B/2},$$

where $(a)_+ = \max(a,0)$. This gives $\|P_{\ell B}\boldsymbol{\epsilon}\|^2 = O_P(K_B(\log n)^{\tau})$.

Furthermore, we have with probability tending to one,

$$\begin{aligned}&\|P_{\ell B}Z_J\boldsymbol{\beta}_J\|^2 = \boldsymbol{\beta}_J'Z_J'P_{\ell B}Z_J\boldsymbol{\beta}_J \\ =&\boldsymbol{\beta}_J'Z_J'H_B\tilde{Z}_{B(\ell)}(\tilde{Z}_{B(\ell)}'H_BH_B\tilde{Z}_{B(\ell)})^{-1}\tilde{Z}_{B(\ell)}'H_BZ_J\boldsymbol{\beta}_J \\ \geq&\lambda_{\min}((\tilde{Z}_{B(\ell)}'H_B\tilde{Z}_{B(\ell)})^{-1})\|\tilde{Z}_{B(\ell)}'H_BZ_J\boldsymbol{\beta}_J\|^2 \\ =&\frac{1}{n\lambda_{\max}(E[\tilde{Z}_k(\mathbf{x}_{1,B(\ell)})H_B\tilde{Z}_k(\mathbf{x}_{1,B(\ell)})'])}\|\tilde{Z}_{B(\ell)}'H_BZ_J\boldsymbol{\beta}_J\|^2.\end{aligned}$$

We are about to find out a lower bound for $\max_{\ell \in J\setminus B}\|\tilde{Z}_{B(\ell)}'H_BZ_J\boldsymbol{\beta}_J\|^2$. Note that equation (4.10) holds for any B and any $\ell \in B^c$, which allows us to write $Y = Z_J\boldsymbol{\beta}_J + \boldsymbol{\gamma}_{k,J} + \boldsymbol{\epsilon}$ into another way. Indeed, since $J = (J \cap B) \cup (J \cap B^c)$, we can write the vector $Z_k(\mathbf{x}_{t,J})$ as $Z_k(\mathbf{x}_{t,J}) \equiv (Z_k(\mathbf{x}_{t,J\cap B})', Z_k(\mathbf{x}_{t,J\cap B^c})')'$, where $Z_k(\mathbf{x}_{t,J\cap B})$ contains all $\mathcal{H}_p(\mathbf{x}_{t,J\cap B})$ for d_1-dimensional ($|J| = d_1$) multiple index p with $|p| = 0, 1, \ldots, k-1$ but whose elements corresponding to $J \cap B^c$ are all zero, whereas $Z_k(\mathbf{x}_{t,J\cap B^c})$ is its complement. As a result, the matrix Z_J can be split into a block matrix $Z_J \equiv [Z_{J\cap B}, Z_{J\cap B^c}]$, and when we split $\boldsymbol{\beta}_J$ conformably as $\boldsymbol{\beta}_J \equiv (\boldsymbol{\beta}_{J\cap B}', \boldsymbol{\beta}_{J\cap B^c}')'$ we have $Y = Z_{J\cap B}\boldsymbol{\beta}_{J\cap B} + Z_{J\cap B^c}\boldsymbol{\beta}_{J\cap B^c} + \boldsymbol{\gamma}_{k,J} + \boldsymbol{\epsilon}$. Now, consider

$$\begin{aligned}\|H_BZ_J\boldsymbol{\beta}_J\|^2 =& \boldsymbol{\beta}_J'Z_J'H_BZ_J\boldsymbol{\beta}_J = (\boldsymbol{\beta}_{J\cap B}'Z_{J\cap B}' + \boldsymbol{\beta}_{J\cap B^c}'Z_{J\cap B^c}')H_BZ_J\boldsymbol{\beta}_J \\ =&\boldsymbol{\beta}_{J\cap B^c}'Z_{J\cap B^c}'H_BZ_J\boldsymbol{\beta}_J = \sum_{\ell \in J\setminus B}\boldsymbol{\beta}_{\ell}\tilde{Z}_{B(\ell)}'H_BZ_J\boldsymbol{\beta}_J \\ \leq& \sum_{\ell \in J\setminus B}\|\boldsymbol{\beta}_{\ell}\|\,\|\tilde{Z}_{B(\ell)}'H_BZ_J\boldsymbol{\beta}_J\| \\ \leq&|J \setminus B|\max_{\ell \in J\setminus B}\|\boldsymbol{\beta}_{\ell}\|\max_{\ell \in J\setminus B}\|\tilde{Z}_{B(\ell)}'H_BZ_J\boldsymbol{\beta}_J\| \\ \leq&d_1\|\boldsymbol{\beta}_J\|\max_{\ell \in J\setminus B}\|\tilde{Z}_{B(\ell)}'H_BZ_J\boldsymbol{\beta}_J\|,\end{aligned}$$

since $\boldsymbol{\beta}_\ell$ is a subvector of $\boldsymbol{\beta}_J$. Further, by (4.3)–(4.5), we obtain that, when k is large,

$$\begin{aligned}\|\boldsymbol{\beta}_J\|^2 =&\sum_{i=0}^{k-1}\sum_{|\mathbf{p}|=i}|a_{i\mathbf{p}}(\boldsymbol{\theta}_1)|^2=\sum_{i=0}^{k-1}c_i^2\sum_{|\mathbf{p}|=i}\binom{i}{\mathbf{p}}(\boldsymbol{\theta}_1)^{2\mathbf{p}}\\ =&\sum_{i=0}^{k-1}c_i^2\|\boldsymbol{\theta}_1\|^2=\sum_{i=0}^{k-1}c_i^2=\|g\|^2(1+o(1)),\end{aligned}$$

by the identification condition and the binomial formula that the power $(\boldsymbol{\theta}_1)^{2\mathbf{p}}$ is the product of each element of $\boldsymbol{\theta}_1$ with corresponding power in $2\mathbf{p}$. Hence,

$$\max_{\ell\in J\backslash B}\|\tilde{Z}'_{B(\ell)}H_BZ_J\boldsymbol{\beta}_J\|^2\geq(d_1\|\boldsymbol{\beta}_J\|)^{-2}\|H_BZ_J\boldsymbol{\beta}_J\|^4.$$

Moreover,

$$\begin{aligned}\|H_BZ_J\boldsymbol{\beta}_J\|^2=&\boldsymbol{\beta}'_JZ'_JH_BZ_J\boldsymbol{\beta}_J=\boldsymbol{\beta}'_{J\cap B^c}Z'_{J\cap B^c}H_BZ_{J\cap B^c}\boldsymbol{\beta}_{J\cap B^c}\\ \geq&\|\boldsymbol{\beta}_{J\cap B^c}\|^2\lambda_{\min}(Z'_{J\cap B^c}H_BZ_{J\cap B^c})\\ =&n\|\boldsymbol{\beta}_{J\cap B^c}\|^2\lambda_{\min}(E[\tilde{Z}_k(\mathbf{x}_{1,B(\ell)})H_B\tilde{Z}_k(\mathbf{x}_{1,B(\ell)})']),\end{aligned}$$

with high probability. We finally have

$$\begin{aligned}\|P_{\ell B}Z_J\boldsymbol{\beta}_J\|^2\geq&\frac{1}{n\lambda_{\max}(\tilde{Z}'_{B(\ell)}H_B\tilde{Z}_{B(\ell)})}\|\tilde{Z}'_{B(\ell)}H_BZ_J\boldsymbol{\beta}_J\|^2\\ \geq&\frac{n\|\boldsymbol{\beta}_{J\cap B^c}\|^4\lambda^2_{\min}(E[\tilde{Z}_k(\mathbf{x}_{1,B(\ell)})H_B\tilde{Z}_k(\mathbf{x}_{1,B(\ell)})']}{(d_1\|\boldsymbol{\beta}_J\|)^2\lambda_{\max}(E[\tilde{Z}_k(\mathbf{x}_{1,B(\ell)})H_B\tilde{Z}_k(\mathbf{x}_{1,B(\ell)})'])}\\ \geq&\frac{nc_0^4c^2(M)}{d_1^2\|g\|^2C(M)k^{3\mu}},\end{aligned}$$

with high probability when k is large, where $c_0=\int g(x)e^{-x^2}dx$ and we suppose $c_0\neq 0$. Otherwise, it can be replaced by any $c_i=\int g(x)h_i(x)e^{-x^2}dx\neq 0$ for some $i\leq k-1$.

It follows from Assumption 3.9 that

$$\max_{\ell\in B^c}(n\widehat{\sigma}^2_B-n\widehat{\sigma}^2_{B(\ell)})\geq\frac{nc_0^4c^2(M)}{d_1^2\|g\|^2C(M)k^{3\mu}},$$

uniformly in B with $|B|\leq M$ with probability tending to one. □

Proof of Corollary 4.1. Note that

$$\begin{aligned}&\mathrm{EBIC}(B)-\mathrm{EBIC}(B(\ell))\\ =&n\log\frac{n\widehat{\sigma}^2_B}{n\widehat{\sigma}^2_{B(\ell)}}+[K_B-K_{B(\ell)}](\log(n)+2\eta\log(d))\\ =&n\log\left[1+\frac{n\widehat{\sigma}^2_B-n\widehat{\sigma}^2_{B(\ell)}}{n\widehat{\sigma}^2_{B(\ell)}}\right]-[K_{B(\ell)}-K_B](\log(n)+2\eta\log(d))\end{aligned}$$

$$\geq n\log\left[1+\frac{n\widehat{\sigma}_B^2-n\widehat{\sigma}_{B(\ell)}^2}{\sum_{i=1}^n(y_i-\bar{y})^2}\right]-k^{|B|+1}(\log(n)+2\eta\log(d))$$
$$\geq\frac{n\widehat{\sigma}_B^2-n\widehat{\sigma}_{B(\ell)}^2}{n^{-1}\sum_{i=1}^n(y_i-\bar{y})^2}-k^M(\log(n)+2\eta\log(d))$$

since $\log(1+x)\geq\min(\log 2,0.5x)$ for any $x>0$, where $\bar{y}$ is the average of $y_i, i=1,\dots,n$.

The conditions of both Corollary 4.1 and Theorem 3.1 ensure that the screening mechanism does not stop when $J\not\subset B$ and $|J\cup B|\leq M$ with high probability. Indeed, if $J\not\subset B_1,\dots,J\not\subset B_m$ and $|J\cup B_m|\leq M$, we have $n^{-1}\sum_{i=1}^n(y_i-\bar{y})^2\geq\widehat{\sigma}_{B_1}^2-\widehat{\sigma}_{B_m}^2$ with probability tending to one. Hence, with high probability we have $\mathrm{Var}(y)>(m-1)D_M$.

If $J\not\subset B_m$ and $m-1>T_M$, then $T_M<\mathrm{Var}(y)/D_M$ which contradicts with the definition of T_M. Hence, at most T_M steps of forward screening we will have J contained in the resultant set. □

Proof of Corollary 4.2. Note that

$$\mathrm{EBIC}(B_m(\ell))-\mathrm{EBIC}(B_m)$$
$$=n\log\frac{n\widehat{\sigma}_{B_m(\ell)}^2}{n\widehat{\sigma}_{B_m}^2}+[K_{B_m(\ell)}-K_{B_m}](\log(n)+2\eta\log(d))$$
$$=n\log\left[1-\frac{n\widehat{\sigma}_{B_m}^2-n\widehat{\sigma}_{B_m(\ell)}^2}{n\widehat{\sigma}_{B_m}^2}\right]+[K_{B_m(\ell)}-K_{B_m}](\log(n)+2\eta\log(d)).$$

Since $J\subset B_m$, similar to the proof of Theorem 4.1 we can apply the proposition of Zhang (Zhang, 2010) to obtain $\widehat{\sigma}_{B_m}^2=E[\epsilon^2]+o_P(1)$. Meanwhile, from a similar derivation to Theorem 4.1, we have

$$n\widehat{\sigma}_{B_m}^2-n\widehat{\sigma}_{B_m(\ell)}^2=\|P_{\ell B}\mathbf{Y}\|^2=O_P(nk^{-s})+O_P(K_{|B_m|}(\log(n))^{\tau}).$$

Thus, $\mathrm{EBIC}(B_m(\ell))-\mathrm{EBIC}(B_m)=K_{B_m(\ell)}(\log(n)+2\eta\log(d))(1+o_P(1))$ and then the scheme of screening stops with probability tending to one. □

SUPPLEMENTARY MATERIAL

Chaohua Dong and Yundong Tu (2023): Supplement to "Semiparametric estimation and variable selection for sparse single index models in increasing dimension," Econometric Theory Supplementary Material. To view, please visit https://doi.org/10.1017/S0266466624000021.

REFERENCES

Ai, C., & Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71, 1795–1843.

Ai, C., & Chen, X. (2007). Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141, 5–43.

Antoniadis, A. (1996). Smoothing noisy data with tapered coiflets series. *Scandinavian Journal of Statistics*, 23, 313–330.

Belloni, A., Chernozhukov, V., Chetverikov, D., & Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186, 345–366.

Belloni, A., Chernozhukov, V., & Wang, L. (2014). Pivotal estimation via square-root lasso in nonparametric regression. *Annals of Statistics*, 42, 757–788.

Chang, J., Chen, S., & Chen, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, 185, 283–304.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3), 759–771.

Chen, X., & Christensen, T. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188, 447–465.

Chen, X., & Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 66, 289–314.

Cheng, M., Honda, T., & Zhang, J. (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, 111(515), 1209–1221.

Cui, X., Hardle, W. K., & Zhu, L. (2011). The EFM approach for single-index models. *Annals of Statistics*, 39, 1658–1688.

Donald, S. G., Imbens, G. W., & Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics*, 152, 28–36.

Dong, C., Gao, J., & Peng, B. (2015). Semiparametric single-index panel data models with cross-sectional dependence. *Journal of Econometrics*, 188, 301–312.

Dong, C., Gao, J., & Tjøstheim, D. (2016). Estimation for single-index and partially linear single-index integrated models. *Annals of Statistics*, 44, 425–453.

Dong, C., Linton, O., & Peng, B. (2021). A weighted sieve estimator for nonparametric time series models with nonstationary variables. *Journal of Econometrics*, 222, 909–932.

Fan, J., Samworth, R. J., & Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research*, 10, 2013–2038.

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

Fan, J., & Liao, Y. (2014). Endogeneity in high dimensions. *Annals of Statistics*, 42, 872–917.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 70(5), 849–911.

Gorst-Rasmussen, A., & Scheike, T. H. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 75(2), 217–245.

Han, X. (2019). Nonparametric screening under conditional strictly convex loss for ultrahigh dimensional sparse data. *Annals of Statistics*, 47(4), 1995–2022.

Hansen, B. E. (2015). *A unified asymptotic distribution theory for parametric and nonparametric least square*. Working paper, University of Wisconsin.

Hardle, W., & Stocker, T. W. (1989). Investigating smooth multiple regression by method of average derivatives. *Journal of the American Statistical Association*, 84, 986–995.

Hardle, W., Hall, P., & Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, 21, 157–178.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1–2), 71–120.

Klein, R. W., & Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 61, 387–421.

Kong, E., Xia, Y., & Zhong, W. (2019). Composite coefficient of determination and its application in ultrahigh dimensional variable screening. *Journal of the American Statistical Association*, 114(528), 1740–1751.

Lv, J., & Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, 37, 3498–3528.

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press

Ma, S., Liang, H., & Tsai, C.-L. (2014). Partially linear single index models for repeated measurements. *Journal of Multivariate Analysis*, 130, 354–375.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79, 147–168.

Pan, W., Wang, X., Xiao, W., & Zhu, H. (2019). A generic sure independence screening procedure. *Journal of the American Statistical Association*, 114(526), 928–937.

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

Power, J. L., Stock, J. H., & Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57, 1403–1430.

Radchenko, P. (2015). High dimensional single index models. *Journal of Multivariate Analysis*, 139, 266–282.

Szego, G. (1975). *Orthogonal Polynomials*. Colloquium Publications XXIII: American Mathematical Association.

Tu, Y., & Wang, S. (2023). Variable screening and model averaging for expectile regressions, *Oxford Bulletin of Economics and Statistics*, 85(3) 574–598.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488), 1512–1524.

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

Xia, Y., Tong, H., Li, W. K., & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction. *Journal of the Royal Statistical Society B*, 64, 363–410.

Yu, Y., & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97, 1042–1054.

Zhang, C. H. (2010). Nearly unbiased variable selection under minmax concave penalty. *Annals of Statistics*, 38, 894–942.

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

Zhang, Y., Lian, H., & Yu, Y. (2020). Ultra-high dimensional single-index quantile regression. *Journal of Machine Learning Research*, 21(224), 1–25.

Zhong, W., Zhu, L., Li, R., & Cui, H. (2016). Regularized quantile regression and robust feature screening for single index models. *Statistica Sinica*, 26(1), 69–95.

https://doi.org/10.1017/S0266466624000021 Published online by Cambridge University Press