A One-Covariate-at-a-Time Multiple Testing Approach to Variable Selection in Additive Models *

Liangjun Su[a], Thomas Tao Yang[b], Yonghui Zhang[c], and Qiankun Zhou[d]

[a] School of Economics and Management, Tsinghua University, China

[b] Research School of Economics, Australian National University, Australia

[c] School of Economics, Renmin University of China, China

[d] Department of Economics, Louisiana State University, USA

February 7, 2024

Abstract

This paper proposes a One-Covariate-at-a-time Multiple Testing (OCMT) approach to choose significant variables in high-dimensional nonparametric additive regression models. Similarly to Chudik, Kapetanios and Pesaran (2018), we consider the statistical significance of individual nonparametric additive components one at a time and take into account the multiple testing nature of the problem. Both one-stage and multiple-stage procedures are considered. The former works well in terms of the true positive rate only if the net effects of all signals are strong enough; the latter helps to pick up hidden signals that have weak net effects. Simulations demonstrate the good finite sample performance of the proposed procedures. As an empirical illustration, we use the OCMT procedure to a dataset extracted from the Longitudinal Survey on Rural Urban Migration in China. We find that our procedure works well in terms of out-of-sample forecast root mean square errors, compared with competing methods such as adaptive group Lasso (AGLASSO).

Keywords: One covariate at a time, multiple testing, model selection, high dimensionality, nonparametric, additive model.

JEL classification: C12, C14, C21, C52.

*We would like to thank the editor, Esfandiar Maasoumi, an associate editor, and two referees for their helpful comments, which have helped improve the paper substantially. Su gratefully acknowledges the support from NSFC under the grant No. 72133002. Zhang gratefully acknowledges the financial support from NSFC (Projects No.71973141 and No.71873033). The computation code used in this paper is available upon request or can be found on the second author's personal website. E-mail: sulj@sem.tsinghua.edu.cn (L. Su), tao.yang@anu.edu.au (T. Yang), yonghui.zhang@ruc.edu.cn (Y. Zhang), and qzhou@lsu.edu (Q. Zhou).

1 Introduction

Variable selection has been playing a pivotal role in econometrics and statistics for statistical learning and scientific discoveries. Notable early contributions include Akaike (1973), Akaike (1974), and Schwartz (1978). These authors suggested a unified approach to model selection, viz., choosing a parameter vector by minimizing the conventional criterion function plus an L_0 penalty to penalize the model size. However, these approaches are not feasible in high-dimensional settings. The seminal work of Tibshirani (1996) addressed this important issue by substituting the L_0 penalty with an L_1 penalty, and it has sparked extensive studies in both statistics and econometrics. Important contributions in the statistics literature include Fan and Li (2001), Zhou and Hastie (2005), Zou (2006), Fan and Lv (2008), Zou and Li (2008), Zhang (2010), Fan, Feng and Song (2011), Fan and Lv (2013), and Fan and Tang (2013). Important contributions in the econometric literature include Belloni et al. (2012), Belloni and Chernozhukov (2013), Belloni, Chernozhukov and Hansen (2014), Belloni et al. (2017), Chernozhukov et al. (2018), and Chudik, Kapetanios and Pesaran (2018). For a comprehensive review, see Fan et al. (2020).

In this paper, we propose a multiple testing approach to variable selection for high-dimensional nonparametric additive models. Recently Chudik, Kapetanios and Pesaran (2018) (CKP hereafter) have proposed a One-Covariate-at-a-time Multiple Testing (OCMT) approach for linear regression models. CKP suggest regressing the dependent variable on each independent variable separately, retaining only those variables that exhibit a high correlation with the dependent variable. This strategy is often referred to as the "*screening approach*". CKP's main contributions are twofold. First, they propose a criterion for variable selection by controlling the probability of choosing all the signals (or with some pseudo-signals) in the model. Second, unlike the usual screening approach that may miss some important "hidden" signals whose net effects on the dependent variable are small, the CKP's OCMT procedure is able to pick up hidden signals with very high probability. The OCMT procedure has been applied in various applications; see, e.g., Kozbur (2020), Chudik, Pesaran and Sharifvaghefi (2021), and Ahmend and Pesaran (2022). In particular, Kozbur (2020) considers a testing-based forward model selection (TBFMS) procedure in linear regression models that inductively selects covariates to add predictive power into a working statistical model. But this latter paper mainly focuses on the error bound and shows that the proposed procedure is able to achieve estimation rates matching those of Least Absolute Shrinkage and Selection Operator (Lasso) and post-Lasso. Furthermore, Sharifvaghefi (2023) extends OCMT to cases with many highly correlated covariates and allows the number of pseudo signals to grow at the same rate as the sample size.

Our paper contributes to the literature by extending the CKP's OCMT approach from parametric models to nonparametric additive models. Like CKP, we estimate the net effect of each variable on the dependent variable one by one, possibly with some preselected variables. The selected variables are those whose net effects exceed some threshold value, specified to ensure the probability of selecting all the signals is very high. The statistics constructed in this paper are much more complicated than the t-statistics in CKP and might not even exhibit a well-behaved limiting distribution. In addition to investigating a different model, our paper differs from that of CKP in some other important aspects. First, we generalize the definition of hidden signals (in Table 2 in Section 2.5). Technical details are updated accordingly. Second, CKP chose tuning parameters similar to a Bonferroni correction of the cumulative distribution function of the standard normal. This choice implicitly requires a certain degree of approximation of the standard normal to the t-statistic distribution in their study. Instead, we select the tuning parameters using the classic Bayesian information criterion (BIC). Third, we add an adaptive group-Lasso-based post-OCMT step to eliminate pseudo-signals that cannot be eliminated with very high probability in CKP. This step adds very little computation burden because the dimension is reduced dramatically before the last-step estimation. This additional step is not needed in theory, but it aims at eliminating the pseudo-signals and thereby enhances the out-of-sample forecasting performance in practice.

One competing method to ours in the literature is the adaptive group Lasso (AGLASSO) proposed by Huang, Horowitz and Wei (2010). The AGLASSO adds some adaptive penalty term to the usual least squares loss function in the spirit of Zou (2006). Even though we also use AGLASSO to eliminate the pseudo-signals after the OCMT procedure, our approach allows much faster computation and provides more reliable estimates than their approach.

We consider various setups in the simulation studies and compare the above post-OCMT AGLASSO procedure with that based on the OCMT alone or the AGLASSO procedure of Huang, Horowitz and Wei (2010) alone. We find that the former one generally outperforms the latter two significantly. We apply our method on a dataset from the Longitudinal Survey on Rural Urban Migration in China (RUMiC) and the empirical results also demonstrate the excellent performance of our procedure in finite samples.

The remainder of the paper is structured as follows. In the next section, we illustrate our approach through a single-stage procedure that is silent to hidden signals. In Section 3, we present the more powerful multiple-stage procedure. We investigate the finite sample properties of our procedure through Monte Carlo experiments in Section 4 and an empirical application in Section 5. We conclude the paper in Section 6. The proofs of all propositions and theorems in the paper are relegated to Appendix B. The online supplement contains some additional technical materials that include the proofs of the technical lemmas in Appendix A and some additional results in the simulation and application. To facilitate reading, We present our procedure in detail for practitioners in Appendix A.1 and the idea of the proofs in Appendices A.2 and A.3.

Notation. For a generic real matrix $\boldsymbol{A} = \{a_{ij}\}$, let $\|\boldsymbol{A}\| = [\lambda_{\max}(\boldsymbol{A}'\boldsymbol{A})]^{1/2}$ denote the spectral norm and $\|\boldsymbol{A}\|_\infty = \max_{ij}|a_{ij}|$. When $\boldsymbol{A}$ is symmetric, $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ denote its maximum and minimum eigenvalues, respectively. For vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|$ denotes its Euclidean norm. For the deterministic series $\{a_n, b_n\}_{n=1}^{\infty}$, we denote $a_n \propto b_n$ if $0 < C_1 \le \liminf_{n\to\infty}|a_n/b_n| \le \limsup_{n\to\infty}|a_n/b_n| \le C_2 < \infty$ for some constants C_1 and C_2, $a_n \lesssim b_n$ if $\limsup_{n\to\infty}|a_n/b_n| \le C < \infty$ for some C that does not depend on n, $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, $a_n \ll b_n$ if $a_n = o(b_n)$, and $a_n \gg b_n$ if $b_n \ll a_n$. $\mathcal{A}^c$ denotes the complement of the set $\mathcal{A}$. $\xrightarrow{P}$ denotes convergence in probability. C and M denote some positive constants that may vary from line to line.

2 The Model and One-Stage Procedure

In this section we give the model and definitions of various versions of signals and noises. Then we will provide the one-stage procedure for variable selection, present the basic assumptions and study the asymptotic properties of our one-stage procedure.

2.1 The Model

Recently, CKP proposed a powerful multiple-stage procedure in order to pick up the hidden signals along with signals with non-zero net effects. In the case where there are no hidden signals, only one stage is needed. Our paper aims to generalize the results of the linear models in CKP to the nonparametric additive models. The notations and technical details in CKP are already quite tedious, and they would be even more so in this paper. To facilitate the exposition, we start with a simple case where there are no pre-determined variables, and we conduct only one-stage multiple testing. We will present the more powerful multiple-stage procedure and show its validity in Section 3.

Suppose the model is

$$Y = f^*(X_1, X_2, \ldots, X_{p^*}) + \varepsilon, \tag{2.1}$$

where Y is the dependent variable, X_1, $X_2, \ldots$, and X_{p^*} are random independent variables, ε is an unobserved error term, and f^* is an unknown smooth function. Even though we have only p^* signal variables, namely X_1, $X_2, \ldots$, X_{p^*}, that should be included into the regression model in (2.1), we do not know this truth before the data reveal the fact. The realistic situation is that the p^* signals are contained in a set $\mathcal{S}_n = \{X_j, j = 1, 2, \ldots, p_n\}$, where p_n can be much larger than the sample size n and $p_n \propto n^{B_p}$ for some $B_p > 0$. $\mathcal{S}_n$ is the set for all candidate variables, and we refer to it as the *active* set. We assume that $E(\varepsilon | X_1, X_2, \ldots, X_{p^*}, X_{p^*+1}, \ldots, X_{p_n}) = 0$. The target of the model selection is to pick up those signals among $\mathcal{S}_n$.

To avoid the curse of dimensionality in nonparametric estimation, we impose the additive structure on f^*, that is,

$$f^*(X_1, X_2, \ldots, X_{p^*}) = \sum_{j=1}^{p^*} f_j^*(X_j). \tag{2.2}$$

We allow the additive components $f_l^*(X_l)$ to change with n but we suppress the dependence of $f_l^*(X_l)$ on n for notional convenience. Obviously, the individual functions f_j^*, $j = 1, \ldots, p^*$, cannot be identified without certain suitable normalizations. We impose the normalization by assuming that $E[f_j^*(X_j)] = 0$ for each j and rewrite the model as

$$Y = \mu + \sum_{j=1}^{p^*} f_j^*(X_j) + \varepsilon, \tag{2.3}$$

where $\mu = E(Y)$. Since μ can be estimated by the sample mean of the Y variable at the usual $\sqrt{n}$-rate and this estimation does not affect the estimation of the nonparametric additive component, for the simplicity of presentation, we assume $\mu = 0$ below.

Since we will focus on the B-spline-based nonparametric theory, it is standard to assume compact support for each regressor (see, e.g., Horowitz and Mammen (2004), Chen (2007) and Huang, Horowitz and Wei (2010)). Without loss of generality, we assume that the support for X_j is $[0,1]$ for all j. Let α_1 be a non-negative integer, $\alpha_2 \in (0,1]$, and $d = \alpha_1 + \alpha_2$. Denote the class of all α_1 times continuously differentiable real-valued functions on $[0,1]$ by $C^{\alpha_1}([0,1])$. Define d-th smooth real-valued functions on $[0,1]$ as

$$\Lambda^d([0,1]) = \left\{ h \in C^{\alpha_1}([0,1]) : \left| h^{(\alpha_1)}(t_1) - h^{(\alpha_1)}(t_2) \right| \le C |t_1 - t_2|^{\alpha_2} \right\}. \tag{2.4}$$

For notational simplicity, we will restrict our attention to the case where f_j's are smooth enough and belong to $\Lambda^d([0,1])$ with $d > 1$. The case of different smoothness parameters only complicates the notation but does not bring in any new insight.

2.2 Signals, Hidden Signals, and Noises

The idea of one-stage procedure is that we estimate the impact of X_l on Y, $l = 1, 2, ..., p_n$, one by one. So we run p_n estimations in total and will keep those variables that are significant enough. Since in each regression we only have one covariate, we are not estimating f_l^* when the explanatory variable is X_l. Instead, we are estimating the conditional expectation $f_l(X_l) \equiv E(Y|X_l)$. We define the *net impact* (or the net effect) of X_l on Y as

$$\theta_l \equiv \left\{ E\left[f_l(X_l)^2 \right] \right\}^{1/2} = \left\{ E\left[\left(\sum_{j=1}^{p^*} \sigma_{lj} \right)^2 \right] \right\}^{1/2},$$

where $\sigma_{lj} = E[f_j^*(X_j)|X_l]$. Since we allow $f_l^*(X_l)$ to change with n, θ_l might change with n as well. But we suppress its dependence on n for notational convenience. Here, σ_{lj} plays the role of the

Table 1: The original definitions of signals and noises from CKP

	$\theta_l \neq 0$	$\theta_l = 0$
$\left\{E[f_l^*(X_l)^2]\right\}^{1/2} \neq 0$	(I) Signals with nonzero net effect	(II) Hidden signals
$\left\{E[f_l^*(X_l)^2]\right\}^{1/2} = 0$	(III) Pseudo-signals	(IV) Noise variables

scaled covariance between X_j and X_l for the linear model considered by CKP. To better understand the connection between the net impact defined here and in CKP, we refer the readers to the results in equations (A.1) and (A.2), where we approximate $f_l(X_l)$ with certain linear functions $f_{nl}(X_l)$. Ignoring the bias from the approximation, one can see the connection more clearly.

Obviously, θ_l can be 0 or close to 0 for signals, and θ_l can be nonzero or large for non-signals. As in CKP, we also have four possibilities as tabulated in Table 1. Cases (I) and (IV) in Table 1 are desirable cases. Case (III) happens when some non-signals are not independent of the signals. The hidden signals defined in Case (II) are rare in the linear case, and it is also rare in the nonparametric case. We generalize the definition of hidden signals to Table 2 in the next section where θ_l is non-zero but small relative to the sample size.

As we shall see, our one-stage procedure is silent on picking up hidden signals and eliminating pseudo-signals. For the hidden signals, we will propose a multiple-stage procedure in Section 3 that can effectively pick them up. To eliminate the pseudo-signals, as a post-procedure we propose to re-estimate the model using adaptive group Lasso in Section 3.4.

We assume that there are p^{**} pseudo-signals. Without loss of generality, we denote them to be

$$\{X_{p^*+1}, X_{p^*+2}, \ldots, X_{p^*+p^{**}}\}.$$

Below we focus on the one-stage procedure in this section and postpone the multi-stage case to the next section.

2.3 The Test Statistic and One-Stage Procedure

Suppose that we have n observations $\{(y_i, x_{1i}, ..., x_{p_n i})\}_{i=1}^n$ that are drawn from the distribution of $(Y, X_1, ..., X_{p_n})$. The data are given in a $n \times (p_n + 1)$ matrix

$$(\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{p_n})$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ and $\boldsymbol{x}_l = (x_{l1}, x_{l2}, \ldots, x_{ln})'$ for $l = 1, ..., p_n$. We propose to choose finite order (e.g., cubic) B-spline basis functions $\{\psi_l(x)\}_{l=1}^{m_n}$ on $[0, 1]$ to approximate the unknown functions f_l's.

B-splines are piecewise-defined polynomial functions that can be used to construct curves and surfaces in numerical analysis. They offer a flexible way to model and control the shape of these curves and surfaces. A B-spline of order n is a piecewise-defined polynomial function of degree $n-1$. B-splines of order one are piecewise constant functions, B-splines of order two are piecewise linear functions, B-splines of order three are piecewise quadratic functions, and so on. The points at which different polynomial pieces connect are called knots, and the knot vector specifies where these knots are. For the detailed definition and properties of B-spline bases, see Stone (1985) and de Boor (2001). We list some properties of the B-spline basis functions in Lemma A.3. Other popular basis functions include polynomials, trigonometric polynomials, splines, and orthogonal wavelets. We refer the readers to Chen (2007) for a nice review about sieve estimation.

Since we normalize $E[f_j^*(X_j)] = 0$, we similarly normalize the basis as

$$\phi_{jl}(x) = \psi_j(x) - n^{-1} \sum_{i=1}^{n} \psi_j(x_{li}).$$

This is a standard practice; see, e.g., Huang, Horowitz and Wei (2010). For notational simplicity, we will write $\phi_j(x)$ for $\phi_{jl}(x)$. Let $P^{m_n}(x) = [\phi_1(x), \phi_2(x), \ldots, \phi_{m_n}(x)]'$, an $m_n \times 1$ vector. Define

$$\boldsymbol{\beta}_l = \left\{ E\left[P^{m_n}(X_l) P^{m_n}(X_l)' \right] \right\}^{-1} E\left[P^{m_n}(X_l) Y \right] \text{ and } U_l = Y - P^{m_n}(X_l)' \boldsymbol{\beta}_l, \tag{2.5}$$

which are the population coefficient and the error term in the regression of Y on $P^{m_n}(X_l)$. Note that we suppress the dependence of $\boldsymbol{\beta}_l$ on the sample size n.

For the one-stage procedure, we conduct the regression of Y on $P^{m_n}(X_l)$, $l = 1, 2, \ldots, p_n$, one by one. Let $\mathbb{X}_{li} = P^{m_n}(x_{li})$ be the approximating function basis at the ith observation for X_l. Let $\mathbb{X}_l = (\mathbb{X}_{l1}, \mathbb{X}_{l2}, \ldots, \mathbb{X}_{ln})'$ be the $n \times m_n$ "design" matrix for X_l. For X_l, we regress $\boldsymbol{y}$ on $\mathbb{X}_l$ to obtain

$$\hat{\boldsymbol{\beta}}_l = \left(\mathbb{X}_l' \mathbb{X}_l \right)^{-1} \mathbb{X}_l' \boldsymbol{y}.$$

We construct the test statistic as[1]

$$\hat{\mathcal{X}}_l = \hat{\boldsymbol{\beta}}_l' \left(\hat{\sigma}_l^{-2} \mathbb{X}_l' \mathbb{X}_l \right) \hat{\boldsymbol{\beta}}_l, \tag{2.6}$$

where $\hat{\sigma}_l^2 = n^{-1} \sum_{i=1}^n \hat{u}_{li}^2$ and $\hat{u}_{li}$ is the residual from the above regression. Then we define the first-stage OCMT selection indictor as

$$\widehat{\mathcal{J}}_l = \mathbf{1}\left(\hat{\mathcal{X}}_l > \varsigma_n \right) \text{ for } l = 1, 2, \ldots, p_n, \tag{2.7}$$

[1] As a referee has noted, one can define an alternative test statistic $\tilde{\mathcal{X}}_l = n\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$, which also works under certain rank conditions (see, e.g., Assumption 10 below). We opted for $\hat{\mathcal{X}}_l$ for two reasons. First, $\hat{\mathcal{X}}_l$ resembles the usual chi-squared statistic under conditional homoskedasticity. Because we do not want to model the conditional heteroskedasticity of unknown form, we cannot take into account conditional heteroskedasticity explicitly in constructing the test statistic $\hat{\mathcal{X}}_l$. Despite this, our asymptotic theory allows for conditional heteroskedasticity in the error term. Second and more importantly, $\hat{\mathcal{X}}_l$ is scale-free whereas $\tilde{\mathcal{X}}_l$ is not. The latter makes it very challenging to choose the range to search the constant C in ς_n defined below.

where $\mathbf{1}(\cdot)$ is the usual indicator function and ς_n is a threshold value.

For the linear model in CKP with one covariate at a time, $\hat{\mathcal{X}}_l$ is asymptotically $\chi^2(1)$ under conditional homoskedasticity, and one can follow their lead to consider threshold values for the associated t-statistics based on the adjusted normal critical values. Nevertheless, such a result is not available in our framework due to the divergent dimension of regressors in the sieve estimation. In addition, the potential presence of conditional heteroskedasticity greatly complicates our analysis too. What we really need is to show that $\hat{\mathcal{X}}_l$ behaves distinctly for signals and noises so that a suitable choice of the threshold value ς_n can help us separate the signals from the noises. For these reasons, we do not associate $\hat{\mathcal{X}}_l$ with any asymptotic distribution. Instead, we will set $\varsigma_n \propto \kappa_n \log(m_n) m_n$ for a positive series κ_n that diverges to infinity slowly as in Assumption 8. For more details, see the remark on Assumption 8 in the next subsection.

2.4 Basic Assumptions

To study the asymptotic properties of the one-stage procedure, we impose the following assumptions.

Assumption 1 $\{y_i, x_{1i}, x_{2i}, \ldots, x_{p_n i}\}_{i=1}^n$ *are independent and identically distributed (i.i.d.) across* i; $E(\varepsilon|X_1, X_2, \ldots, X_{p^*}, X_{p^*+1}, \ldots, X_{p_n}) = 0$.

Assumption 2 p^* *is a positive integer that does not vary with* n. $p^{**} \lesssim n^{B_{p^{**}}}$ *and* $p_n \propto n^{B_p}$ *for some* $B_p > B_{p^{**}} \geq 0$.

Assumption 3 *The support for* X_l *is* $[0,1]$, $l = 1, \ldots, p_n$. *The density function for* X_l *is bounded and bounded away from* 0.

Assumption 4 $\Pr(|\varepsilon| > t) \leq C_1 \exp(-C_2 t^s)$ *holds for all* $t > 0$ *and some* s, C_1, $C_2 > 0$.

Assumption 5 $f_l(\cdot) = E(Y|X_l = \cdot) \in \Lambda^d([0,1])$ *with* $d > 1$ *for* $l = 1, \ldots, p_n$.

Assumption 6 $\left|f_j^*(\cdot)\right|$, $j = 1, ..., p^*$, *are uniformly bounded.* $E\left(\varepsilon^2|X_1, X_2, ..., X_{p_n}\right)$ *is uniformly bounded almost surely.*

Assumption 7 $m_n \propto n^{B_m}$ *with* $1/(1+2d) < B_m < 1/3$.

Assumption 8 $\varsigma_n \propto \kappa_n \log(m_n) m_n$ *for some* κ_n *such that* $\kappa_n > 0$, $\kappa_n \to \infty$, *and* $\kappa_n = O\left[(\log n)^{\epsilon}\right]$ *for some small positive* ϵ, *as* $n \to \infty$.

Assumption 1 imposes an i.i.d. condition on the observations and a standard conditional moment restriction. The extension to weakly dependent observations is possible but left for future research.

The generalization to independently non-identically distributed (i.n.i.d.) case is straightforward because the main inequalities in Lemmas A.1 and A.2 allow for i.n.i.d. observations. We keep using the i.i.d. assumption for notational convenience. In Assumption 2, we assume that the number of signals is fixed; the number of pseudo-signals is allowed to increase as n increases but at a slower rate than that of the total candidate variables. It is also possible to allow p^* to diverge to infinity (see Section 3.5). Assumption 3 restricts the support of X to be $[0,1]$. This is a very common condition for nonparametric additive models; see, e.g., Li (2000) and Horowitz and Mammen (2004). Assumption 4 imposes some tail conditions ε, which is also assumed in CKP but weaker than the commonly used sub-exponential condition in the variable selection literature and the one for the adaptive group Lasso in Huang, Horowitz and Wei (2010). The primary use of this condition is to derive *probability* bound for errors.

Assumption 5 imposes fairly weak smooth condition on $f_l(x)$, which is weaker than the commonly used condition $d \geq 2$. Note that $d > 1$ is needed for Assumption 7. Assumption 6 is a technical assumption needed to simplify the proof. Specifically, the boundedness of f_j^* implies U_l defined in equation (A.1) has the same tail behavior as ε. The bounded conditional second moment of ε is to ensure some nice properties of $U_l \phi_j(X_l)$ and this assumption is also common in the sieve literature (see, e.g., Newey (1997)). This assumption is mild given that we assume all X_l's have compact support and the tails of ε decay exponentially fast. Assumption 7 imposes conditions on m_n. First, we need $B_m < 1/3$ such that $nm_n^{-3} \to \infty$. The last condition is necessary for $\left\| n^{-1}\mathbb{X}_l'\mathbb{X}_l \right\| \propto m_n^{-1}$ to hold with very high probability. To see why, note that Lemma A.3 in Appendix A.4 suggests that $\left\| E\left[P^{m_n}(X_l)\, P^{m_n}(X_l)'\right] \right\| \propto m_n^{-1}$. We need $m_n^{-1} \gg (m_n/n)^{1/2}$, or equivalently, $nm_n^{-3} \to \infty$, in order to ensure that $n^{-1}\mathbb{X}_l'\mathbb{X}_l$ is close to $E\left[P^{m_n}(X_l)\, P^{m_n}(X_l)'\right]$ with very high probability. Second, we need $B_m > 1/(1+2d)$ to ensure that the approximation bias is asymptotically negligible in comparison with the asymptotic variance term: $m_n^{-d} \ll (m_n/n)^{1/2}$.

Assumption 8 imposes conditions on the threshold value ς_n that ensures the separability of the signals from noises. If the true value of $\boldsymbol{\beta}_l$ is $\mathbf{0}$, $\hat{\mathcal{X}}_l$ in equation (2.6) is $O_P(m_n)$. In this case, to ensure $\widehat{\mathcal{J}}_l = 0$ with very high probability, we can take $\varsigma_n \propto \kappa_n \log(m_n)\, m_n$ and lose some power up to $\kappa_n \log(m_n)$. Here, κ_n can be any series diverging to infinity slowly, e.g., $[\log(m_n)]^{\epsilon}$ for some small $\epsilon > 0$. The loss of the power to some degree is inevitable because of the nature of the multiple testing procedure when the number of tests goes to infinity. In contrast, CKP choose their threshold by the Bonferroni correction of the standard normal. We choose not to follow them because of the following reasons. First, given the divergence of m_n, $\hat{\mathcal{X}}_l$ does not converge to a chi-square distribution asymptotically even in the homoskedastic case so that we cannot use chi-square distribution to approximate the finite sample distribution of $\hat{\mathcal{X}}_l$. Under conditional heteroskedasticity, $\hat{\mathcal{X}}_l$ does not converge to a chi-square distribution even if m_n is held fixed. So our procedure does not rely on the chi-square approximation. Second, even if we can do the approximation, the cumulative density function (CDF) of a chi-square distribution is very complicated. We do not have a rate for the inverse of its CDF evaluated at a certain rate (e.g., n^{-C}) like the case of normal CDF. For these reasons, we do our selection based on the asymptotic results. Specifically, we will take $\varsigma_n = C\kappa_n \log(m_n) m_n$ for some κ_n and choose the value of C by the classic BIC. The details can be found in Appendix A.1. This ς_n helps separate the signals from the noises with very high probability, as demonstrated in the next section. Recall that in Assumption 2 we assume that p_n go to infinity at a polynomial rate of n, same as m_n. This ensures that $\log(m_n) \propto \log p_n$. Therefore, theoretically we only need to put $\log(m_n)$ in ς_n to have a control of p^{**} and p_n for the TPR, FDR and FPR defined after Proposition 2.1 below. We postpone the discussion on the comparison of technical conditions required for our procedure and those required for the AGLASSO to Section 3.2.

2.5 The Asymptotic Properties of $\hat{\mathcal{X}}_l$ and the One-Stage Procedure

We present the first theoretical result in this paper. It derives the probability bounds for the "Type-I" and "Type-II" errors.

Proposition 2.1 *Suppose that Y is given by equation (2.1) and Assumptions $1-8$ hold.*

(i) If $\theta_l \lesssim \log(m_n)^{1/2}(m_n/n)^{1/2}$, then for sufficiently large n we have

$$\Pr\left(\hat{\mathcal{X}}_l \geq \varsigma_n\right) \leq \exp\left(-C_1 m_n^{-1}\varsigma_n + \log m_n\right) + C_2\exp\left(-C_3 n^{C_4}\right)$$
$$\leq n^{-M} + C_2\exp\left(-C_3 n^{C_4}\right)$$

for any fixed constant $M > 0$ and some positive constants C_1, C_2, C_3, and C_4.

(ii) If $\theta_l \gtrsim \kappa_n \log(m_n)^{1/2}(m_n/n)^{1/2}$ with κ_n specified in Assumption 8, then for sufficiently large n we have

$$\Pr\left(\hat{\mathcal{X}}_l \geq \varsigma_n\right) \geq 1 - n^{-M} - C_5\exp\left(-C_6 n^{C_7}\right)$$

for any fixed constant $M > 0$ and some positive constants C_5, C_6, and C_7.

The proof of Proposition 2.1 is tedious. We provide some technical details in Appendix A.2 before we formally prove the proposition in Appendix B. An implication of Proposition 2.1 is that for the well-chosen threshold value ς_n, the above one-stage procedure can separate the signals with $\theta_l \gtrsim \kappa_n \log(m_n)^{1/2}(m_n/n)^{1/2}$ from the noises with $\theta_l = 0$. Of course, we may have some intermediate case where $0 < \theta_l \lesssim \log(m_n)^{1/2}(m_n/n)^{1/2}$ for which the above procedure fails to do so. Note that the convergence rate for each additive term in Stone (1985) is $(m_n/n)^{1/2}$. Therefore, our procedure loses some power up to the order of $\log(n)$, a common scenario in the variable selection literature.

Table 2: The generalized definitions of signals and noises

	$\theta_l \gg \log(m_n)^{1/2}(m_n/n)^{1/2}$	$\theta_l \lesssim \log(m_n)^{1/2}(m_n/n)^{1/2}$
$\left\{E[f_l^*(X_l)^2]\right\}^{1/2} \neq 0$	(I) Signals	(II) Hidden signals
$\left\{E[f_l^*(X_l)^2]\right\}^{1/2} = 0$	(III) Pseudo-signals	(IV) Noise variables

Following the literature and CKP, we define the true positive rates (TPR) and the false positive rates (FPR) respectively as

$$\mathrm{TPR}_n = \frac{\sum_{l=1}^{p_n} \mathbf{1}\left(\widehat{\mathcal{J}}_l = 1 \text{ and } \left\{E\left[f_l^*(X_l)^2\right]\right\}^{1/2} \neq 0\right)}{\sum_{l=1}^{p_n} \mathbf{1}\left(\left\{E\left[f_l^*(X_l)^2\right]\right\}^{1/2} \neq 0\right)}, \text{ and}$$

$$\mathrm{FPR}_n = \frac{\sum_{l=1}^{p_n} \mathbf{1}\left(\widehat{\mathcal{J}}_l = 1 \text{ and } \left\{E\left[f_l^*(X_l)^2\right]\right\}^{1/2} = 0\right)}{\sum_{l=1}^{p_n} \mathbf{1}\left(\left\{E\left[f_l^*(X_l)^2\right]\right\}^{1/2} = 0\right)}.$$

Based on the test statistic defined in equation (2.6) and its property developed in Proposition 2.1, we introduce the generalized definitions of signals and noises in Table 2. With this definition of hidden signals, our result in Section 3 provides theoretical justification for the necessity of a multi-stage procedure capable of detecting signals not identified in the first stage, yet with non-zero net effects.

With a little abuse of notation, we continue to use p^* and p^{**} to denote the number of signals and pseudo-signals at the sample level, and assume that they satisfy Assumption 2. Adopting the definitions in Table 2, we define the false discovery rates (FDR) as

$$\mathrm{FDR}_n = \frac{\sum_{l=1}^{p_n} \mathbf{1}\left(\widehat{\mathcal{J}}_l = 1,\ \left\{E\left[f_l^*(X_l)^2\right]\right\}^{1/2} = 0, \text{ and } \theta_l \lesssim \log(m_n)^{1/2}(m_n/n)^{1/2}\right)}{\sum_{l=1}^{p_n} \widehat{\mathcal{J}}_l + 1}.$$

Apparently, the definitions in Table 2 generalize the definitions in Table 1.

The one-stage procedure is valid in terms of TPR, only if the net effects of all signals are strong enough. Specifically, we need the following assumption.

Assumption 9 *There are no hidden signals. That is,* $\theta_j \gtrsim \kappa_n \log(m_n)^{1/2}(m_n/n)^{1/2}$ *for some slowly divergent series* κ_n *as in Assumption 8 for all* $j = 1, 2, ..., p^*$.

We note the definition of no hidden signals in the above assumption is equivalent to the one in Table 2, given the way κ_n is defined in Assumption 8. If Assumption 9 fails to hold, we need the multiple-stage procedure to pick up the hidden signals.

Based on the results in Proposition 2.1, we present the results for TPR_n, FPR_n, and FDR_n in the following theorem.

Theorem 2.1 *Suppose that Y is given by equation (2.1) and Assumptions 1 − 9 hold. Then after some large n,*

(i) $E(\mathrm{TPR}_n) \geq 1 - C_1 \exp\left(-C_2 n^{C_3}\right)$ *for some positive constants* $C_1, C_2,$ *and* C_3;

(ii) $E(\mathrm{FPR}_n) \leq p^{**}/(p_n - p^*) + C_4 n^{-M} + C_5 \exp\left(-C_6 n^{C_7}\right)$ *for any fixed positive large constant* M *and some positive constants* C_4, $C_5, C_6,$ *and* C_7;

(iii) $\mathrm{FDR}_n \xrightarrow{P} 0$.

Theorem 2.1 implies that all of TPR_n, FPR_n and FDR_n can be well controlled provided we assume away hidden signals at the sample level. Note that Theorem 2.1(i)–(ii) focuses on the asymptotic properties of TPR_n and FPR_n while the last part of Theorem 2.1 reveals that the false discovery rate is asymptotically vanishing in large samples.

In the next section, we turn to the multiple-stage procedure that does not rely on Assumption 9.

3 The Multiple-Stage Procedure

In this section we propose a multiple-stage procedure to select variables for the nonparametric additive models.

3.1 The Test Statistic with Pre-selected Variables

As mentioned above, we may not identify a signal X_l whose net effect satisfies $\theta_l \lesssim \kappa_n \log(m_n)^{1/2} (m_n/n)^{1/2}$, even in the case where the *marginal* effect of X_l on Y, namely, $\left\{E\left[f_l^*(X_l)^2\right]\right\}^{1/2}$, is large enough.[2] Consequently, Theorem 2.1 does not hold without Assumption 9 which assumes away small sample hidden signals. In contrast, Lemma A.7 in Appendix A.4 underpins the result that as long as $\left\{E\left[f_l^*(X_l)^2\right]\right\}^{1/2} \gtrsim \kappa_n \log(m_n)^{1/2} (m_n/n)^{1/2}$ for some slowly divergent series κ_n and some full rank condition holds, the net effect of X_l will be strong enough to be picked up at certain stage of a multiple-stage procedure. A by-product of this Lemma is the existence of at least one signal for Stage 1.

To introduce the multiple-stage procedure, we need some extra notations. Suppose at certain stage after stage 1, we have pre-selected ι_n variables from the active set $\mathcal{S}_n = \{X_j,\ j = 1, ..., p_n\}$ based on some selection procedure to be described below. To avoid confusion, we denote these pre-selected variables as $Z_1, \ldots, Z_{\iota_n}$. Let $\boldsymbol{Z} \equiv (Z_1, \ldots, Z_{\iota_n})'$ and $P^{m_n}(\boldsymbol{Z}) = (P^{m_n}(Z_1)', \ldots,\ P^{m_n}(Z_{\iota_n})')'$. Note that $P^{m_n}(\boldsymbol{Z})$ is an $\iota_n m_n \times 1$ vector for $\boldsymbol{Z}$. At the next stage, we consider the nonparametric additive

[2]The marginal effect defined here is slightly different from that in the econometrics literature. For example, the marginal effect of X_l on Y is defined as β_l in the CKP's linear model: $Y = \beta_0 + \sum_{l=1}^{p^*} X_l \beta_l + \varepsilon$, but it refers to $X_l \beta_l$.in this paper.

regression of Y on $\boldsymbol{Z}$ and an X_l that has not been selected so far and we do this one by one for all $p_n - \iota_n$ non-selected variables X_l. We define the impact of X_l on Y after controlling $\boldsymbol{Z}$ as

$$\begin{aligned}
\theta_{l,\boldsymbol{Z}} &\equiv \left\{E\left\{E\left[Y - P^{m_n}(\boldsymbol{Z})' \Phi_{\boldsymbol{Z}}^{-1} E\left[P^{m_n}(\boldsymbol{Z}) Y\right] \middle| X_l\right]\right\}^2\right\}^{1/2} \\
&= \left\{E\left\{\sum_{j=1}^{p^*} E\left[f_j^*(X_j) - P^{m_n}(\boldsymbol{Z})' \Phi_{\boldsymbol{Z}}^{-1} E\left[P^{m_n}(\boldsymbol{Z}) f_j^*(X_j)\right] \middle| X_l\right]\right\}^2\right\}^{1/2} \\
&= \left\{E\left[\left(\sum_{j=1}^{p^*} \mu_{lj,\boldsymbol{Z}}\right)^2\right]\right\}^{1/2},
\end{aligned}$$

where $\Phi_{\boldsymbol{Z}} \equiv E[P^{m_n}(\boldsymbol{Z}) P^{m_n}(\boldsymbol{Z})']$ and $\mu_{lj,\boldsymbol{Z}} \equiv E\left\{f_j^*(X_j) - P^{m_n}(\boldsymbol{Z})' \Phi_{\boldsymbol{Z}}^{-1} E\left[P^{m_n}(\boldsymbol{Z}) f_j^*(X_j)\right] \middle| X_l\right\}$ denotes the effect of X_l on $f_j^*(X_j)$ after controlling the effects of $\boldsymbol{Z}$. Apparently, we suppress the dependence of $\theta_{l,\boldsymbol{Z}}$ on the sample size n.

At the sample level, let $\mathbb{Z}$ and $\mathbb{X}_l$ denote the $n \times m_n \iota_n$ and $n \times m_n$ "design matrices" for $\boldsymbol{Z}$ and X_l, respectively. That is,

$$\mathbb{Z} \equiv (\mathbb{Z}_1, \ldots, \mathbb{Z}_{\iota_n}) \text{ and } \mathbb{X}_l \equiv (\mathbb{X}_{l1}, \ldots, \mathbb{X}_{ln})', \tag{3.1}$$

where $\mathbb{Z}_l = (\mathbb{Z}_{l1}, \ldots, \mathbb{Z}_{ln})'$ is a $n \times m_n$ matrix, $\mathbb{Z}_{li} = P^{m_n}(z_{li})$ and $\mathbb{X}_{li} = P^{m_n}(x_{li})$. Define $M_{\mathbb{Z}} \equiv I_n - \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'$. By the result of partitioned regressions, the coefficient of $P^{m_n}(X_l)$ is estimated by

$$\hat{\boldsymbol{\beta}}_{l,\boldsymbol{Z}} = \left(\mathbb{X}_l' M_{\mathbb{Z}} \mathbb{X}_l\right)^{-1} \mathbb{X}_l' M_{\mathbb{Z}} \boldsymbol{y}.$$

To determine whether X_l should be treated as a signal variable, we propose the following test statistic

$$\hat{\mathcal{X}}_{l,\boldsymbol{Z}} = \hat{\boldsymbol{\beta}}_{l,\boldsymbol{Z}}' \left(\hat{\sigma}_{l,\boldsymbol{Z}}^{-2} \mathbb{X}_l' M_{\mathbb{Z}} \mathbb{X}_l\right) \hat{\boldsymbol{\beta}}_{l,\boldsymbol{Z}} = \left(\boldsymbol{y}' M_{\mathbb{Z}} \mathbb{X}_l\right) \left(\hat{\sigma}_{l,\boldsymbol{Z}}^2 \mathbb{X}_l' M_{\mathbb{Z}} \mathbb{X}_l\right)^{-1} \left(\mathbb{X}_l' M_{\mathbb{Z}} \boldsymbol{y}\right) \tag{3.2}$$

where $\hat{\sigma}_{l,\boldsymbol{Z}}^2 = n^{-1} \sum_{i=1}^n \hat{\varepsilon}_{li}^2$ and $\hat{\varepsilon}_{li}$ is the residual from the regression $\boldsymbol{y}$ on $(\mathbb{Z}, \mathbb{X}_l)$.

We will study the asymptotic properties of $\hat{\mathcal{X}}_{l,\boldsymbol{Z}}$ in the next subsection which lay down the foundation for our multiple-stage procedure.

3.2 The Asymptotic Properties of $\hat{\mathcal{X}}_{l,\boldsymbol{Z}}$

To study the asymptotic properties of $\hat{\mathcal{X}}_{l,\boldsymbol{Z}}$, we impose the following technical conditions.

Assumption 2' p^* is a positive integer that does not vary with n. $p^{**} \lesssim n^{B_{p^{**}}}$ and $p_n \propto n^{B_p}$ for some $B_p > B_{p^{**}} \geq 0$. Further, $B_{p^{**}} < (1 - 3B_m)/2$.

Assumption 5' $f_l(\cdot) = E(Y|X_l = \cdot) \in \Lambda^d([0,1])$ with $d > 1$ for $l = 1, \ldots, p_n$. $f_j^* \in \Lambda^d([0,1])$ with $d > 1$ for $j = 1, \ldots, p^*$.

Assumption 9' $\left\{E[f_j^*(X_j)^2]\right\}^{1/2} \gtrsim \kappa_n \log(m_n)^{1/2} (m_n/n)^{1/2}$ for some slowly divergent series κ_n as in Assumption 8 and for $j = 1, \ldots, p^*$.

Assumption 10 *Let* $\boldsymbol{X}_1^{p^*+p^{**}} \equiv (X_1, \ldots, X_{p^*}, X_{p^*+1}, \ldots, X_{p^*+p^{**}})'$, *the vector of all signals and pseudo signals. Similarly, let* $P^{m_n}(\boldsymbol{X}_1^{p^*+p^{**}}) \equiv \left[P^{m_n}(X_1)', \ldots, P^{m_n}(X_{p^*+p^{**}})'\right]'$ *and* $\Phi_{\boldsymbol{X}_1^{p^*+p^{**}}} \equiv E[P^{m_n}(\boldsymbol{X}_1^{p^*+p^{**}})\ P^{m_n}(\boldsymbol{X}_1^{p^*+p^{**}})']$. *Assume that*

$$B_{X1} m_n^{-1} \leq \lambda_{\min}\left(\Phi_{\boldsymbol{X}_1^{p^*+p^{**}}}\right) \leq \lambda_{\max}\left(\Phi_{\boldsymbol{X}_1^{p^*+p^{**}}}\right) \leq B_{X2} m_n^{-1}$$

holds for some positive constants B_{X1} *and* B_{X2}, *and it also holds when* $\boldsymbol{X}_1^{p^*+p^{**}}$ *is augmented by an arbitrary element* X_l *with* $l > p^* + p^{**}$.

Assumption 11 *If the pre-selected variables* $\boldsymbol{Z}$ *are either signals or pseudo-signals, the effect of noise variable defined in Table 2 on* Y *is still weak. That is,* $\theta_{l,\boldsymbol{Z}} \lesssim \log(m_n)^{1/2} (m_n/n)^{1/2}$ *for all* $l = p^* + p^{**} + 1, \ldots, p_n$, *when all variables in* $\boldsymbol{Z}$ *are either signals or pseudo-signals.*

Assumption 2' strengthens Assumption 2 by adding one more condition on $B_{p^{**}}$. It is imposed to ensure the good property of $\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}$ defined in equation (A.11) (see Lemma A.10). This condition can be very restrictive on $B_{p^{**}}$. For example, when $B_m = 1/4$, this condition implies $B_{p^{**}} < 1/8$. Assumption 5' strengthens Assumption 5 so that equation (A.4) in Appendix A.2 holds. As remarked at the beginning of last subsection, we do not impose Assumption 9 for the multiple-stage procedure, as long as the marginal effect of the signal is not too weak as imposed in Assumption 9'. This assumption seems inevitable. It is analogous to Assumption 6 in CKP for the linear regression model and similar to the so-called 'beta-min' condition that is commonly assumed in the penalized regression literature (see, e.g., Chapter 7.4 of Buhlmann and van de Geer (2011)). Assumption 10 is the common rank condition for nonparametric additive regressions. We also require it to hold for the case when we add one noise variable. This seems inevitable because we will run the regression with regressors being all signals and pseudo-signals plus one noise variable in the multiple-stage procedure with very high probability. Assumption 11 is also inevitable and implicitly imposed in CKP for the linear regression models.

It is worth mentioning that Assumption 10 plays a similar role to the "restrictive eigenvalues" condition in Bickel, Ritov and Tsybakov (2009) and Belloni et al. (2012). The restrictive-eigenvalues condition requires certain full rank conditions on all possible design matrices composed of a certain number of covariates. In contrast, our OCMT only requires full rank conditions on the design matrices composed of signals and pseudo-signals, and permits arbitrary correlations among the noise variables. Obviously, these two sets of conditions are non-nested. In addition, Huang, Horowitz and Wei (2010) impose essentially the same set of assumptions except the rank conditions discussed here. The main advantage of the OCMT is that it does not require any numerical min-search of an objective function, can be computed much faster, and deliver more reliable results.

The following proposition presents the probability bounds for the "Type-I" and "Type-II" errors when we have some pre-selected variables.

Proposition 3.1 *Suppose that Y is given by equation (2.1), Assumptions 1, 2', 3, 4, 5', 6, 7, 8, and 10 hold, and the pre-selected variables $\boldsymbol{Z}$ are either signals or pseudo-signals.*

(i) If $\theta_{l,\boldsymbol{Z}} \lesssim \log(m_n)^{1/2} m_n^{1/2} n^{-1/2}$, then

$$\Pr\left(\hat{\mathcal{X}}_{l,\boldsymbol{Z}} \geq \varsigma_n\right) \leq \exp\left(-C_1 m_n^{-1} \varsigma_n + \log m_n\right) + C_2 \exp\left(-C_3 n^{C_4}\right) \leq n^{-M} + C_2 \exp\left(-C_3 n^{C_4}\right) \tag{3.3}$$

for any fixed large constant $M > 0$ and some positive constants $C_1, C_2, C_3,$ and C_4, after some large n.

(ii) If $\theta_{l,\boldsymbol{Z}} \gtrsim \kappa_n \log(m_n)^{1/2} m_n^{1/2} n^{-1/2}$ with κ_n specified in Assumption 8, then

$$\Pr\left(\hat{\mathcal{X}}_{l,\boldsymbol{Z}} \geq \varsigma_n\right) \geq 1 - n^{-M} - C_5 \exp\left(-C_6 n^{C_7}\right)$$

for any fixed large constant $M > 0$ and some positive constants $C_5, C_6,$ and C_7, after some large n.

The proof of Proposition 3.1 is rather tedious. We provide some technical discussions on the proof in Appendix A.3 before we formally prove it in Appendix B.

Like Proposition 2.1, Proposition 3.1 implies that for the well-chosen threshold value ς_n, the use of the test statistic $\hat{\mathcal{X}}_{l,\boldsymbol{Z}}$ helps to separate variables with large value of $\theta_{l,\boldsymbol{Z}}$ from those with small value of $\theta_{l,\boldsymbol{Z}}$. This observation will be used in our multiple-stage procedure to select all signal variables.

3.3 The Multiple-Stage Procedure

We present the multiple-stage procedure as follows.

We conduct the first-stage selection as in Section 2.3 by constructing the test statistic $\hat{\mathcal{X}}_l$ as in equation (2.6) and using the threshold value ς_n that satisfies the condition in Assumption 8. We re-label the selection indicator in equation (2.7) as

$$\widehat{\mathcal{J}}_{l,(1)} = \mathbf{1}\left(\hat{\mathcal{X}}_l > \varsigma_n\right) \text{ for } l = 1, 2, \ldots, p_n.$$

We collect all the variables selected in stage 1 into the vector $\boldsymbol{Z}_{(1)}$, and denote the index set of the selected variables by $S_{(1)}$. For the second stage, we denote the index set of the active variables in stage 2 by $\Psi_{(2)}$ where $\Psi_{(2)} = \{1, 2, \ldots, p_n\} \backslash S_{(1)}$. In the second stage, we regress Y on $P^{m_n}(X_l)$ with $P^{m_n}\left(\boldsymbol{Z}_{(1)}\right)$ as pre-selected variables one by one for $l \in \Psi_{(2)}$. We construct the test statistic $\hat{\mathcal{X}}_{l,\boldsymbol{Z}_{(1)}}$ as in equation (3.2). We select the variable X_l if $\widehat{\mathcal{J}}_{l,(2)} = 1$, where

$$\widehat{\mathcal{J}}_{l,(2)} = \mathbf{1}\left(\hat{\mathcal{X}}_{l,\boldsymbol{Z}_{(1)}} > \varsigma_n\right) \text{ for } l \in \Psi_{(2)}.$$

We add all the variables selected in stage 2 into the set of variables selected in stage 1 as a new vector, and we denote it as $\boldsymbol{Z}_{(2)}$. We denote the index set of the selected variables ($\boldsymbol{Z}_{(2)}$) by $S_{(2)}$ and the index set of the active variables for stage 3 as $\Psi_{(3)}$, where $\Psi_{(3)} = \{1, 2, \ldots, p_n\} \backslash S_{(2)}$. And so on and so forth. For stage k, we denote the pre-selected variables as $\boldsymbol{Z}_{(k-1)}$, and the index set of the active variables as $\Psi_{(k)}$. Then we regress Y on $P^{m_n}(X_l)$ with $P^{m_n}\left(\boldsymbol{Z}_{(k-1)}\right)$ as pre-selected variables one by one for $l \in \Psi_{(k)}$. We construct the test statistic $\hat{\mathcal{X}}_{l,\boldsymbol{Z}_{(k-1)}}$ as in equation (3.2). We select the variable X_l if $\widehat{\mathcal{J}}_{l,(k)} = 1$, where

$$\widehat{\mathcal{J}}_{l,(k)} = \mathbf{1}\left(\hat{\mathcal{X}}_{l,\boldsymbol{Z}_{(k-1)}} > \varsigma_n\right) \text{ for } l \in \Psi_{(k)}.$$

We add all the variables selected in stage k into the set of variables selected in stage $k-1$ as a new vector, and we denote it as $\boldsymbol{Z}_{(k)}$. We stop the procedure at a stage in which no new variables are selected. We denote the stage, in which one or more variables are selected but no new variables are selected after that, as $\hat{k}_s$. So the OCMT procedures stops after stage $\hat{k}_s$. The selection indicator for variable X_l of the OCMT procedure is defined as follows

$$\widehat{\mathcal{J}}_l = \sum_{k=1}^{\hat{k}_s} \widehat{\mathcal{J}}_{l,(k)}. \tag{3.4}$$

By construction, $\widehat{\mathcal{J}}_l$ is either 1 or 0. It takes value 1 if X_l is selected in the OCMT procedure and 0 otherwise.

The following theorem mainly studies the asymptotic properties of the multiple-stage procedure in terms of TPR, FPR and FDR.

Theorem 3.1 *Suppose that Assumptions 1, 2', 3, 4, 5', 6, 7, 8, 9', 10, and 11 hold. Then after some large n,*

(i) $\Pr\left(\hat{k}_s > p^*\right) \leq n^{-M_6} + C_{19}\exp\left(-C_{20}n^{C_{21}}\right)$ *for some fixed large positive number M_6 and some positive constants C_{19}, C_{20}, and C_{21};*

(ii) $E(\mathrm{TPR}_n) \geq 1 - C_1\exp\left(-C_2 n^{C_3}\right)$ *for some positive constants C_1, C_2, and C_3;*

(iii) $E(\mathrm{FPR}_n) \leq p^{**}/(p_n - p^*) + C_4 n^{-M}$ *for some positive C_4 and any fixed positive large constant M;*

(iv) $\mathrm{FDR}_n \xrightarrow{P} 0$.

Theorem 3.1(i) implies the OCMT procedure can terminate at step p^* with very high probability. Theorem 3.1(ii)-(iv) implies that all of TPR_n, FPR_n and FDR_n can be well controlled. Of course, when $n \to \infty$, we need to conduct at most p^*-stage procedure to determine all signals and eliminate all noise variables, with very high probability.

3.4 Dealing with Pseudo-signals

Because of the nature of the OCMT procedure, pseudo-signals cannot be excluded from the selection list with high probability. To eliminate pseudo-signals, we propose to employ the adaptive group Lasso after the OCMT. Then by the properties of the adaptive group Lasso, the event that the pseudo-signals are excluded and the signals are kept occurs with probability approaching one (w.p.a.1). We present this result in Theorem 3.2 below. We denote the set of the variables selected by the OCMT procedure as $\hat{S}_{\text{OCMT}} \equiv S_{(\hat{k}_s)}$, collect them into a vector $\boldsymbol{Z}_{\text{OCMT}}$, and denote its dimension as $\hat{p}_{\text{OCMT}}$. Similarly, let $\mathbb{Z}_{\text{OCMT}}$ $(n \times \hat{p}_{\text{OCMT}} m_n)$ denote the design matrix for $\boldsymbol{Z}_{\text{OCMT}}$ as in equation (3.1). The post-OCMT adaptive group Lasso procedure goes as follows:

1. Obtain the group Lasso estimator by searching $\boldsymbol{\beta}_n$ $(\hat{p}_{\text{OCMT}} m_n \times 1)$ to minimize

$$L_{n1}(\boldsymbol{\beta}_n, \lambda_{n1}) = \|\boldsymbol{y} - \mathbb{Z}_{\text{OCMT}} \boldsymbol{\beta}_n\|^2 + \lambda_{n1} \sum_{j=1}^{\hat{p}_{\text{OCMT}}} \left\|\boldsymbol{\beta}_{nj}\right\|,$$

   where λ_{n1} is a positive tuning parameter, $\boldsymbol{\beta}_n = (\boldsymbol{\beta}'_{n1}, ..., \boldsymbol{\beta}'_{n\hat{p}_{\text{OCMT}}})'$, $\boldsymbol{\beta}_{nj}$ is a $m_n \times 1$ vector of coefficients of the B-spline basis for the j-th element in $\boldsymbol{Z}_{\text{OCMT}}$. Denote the above estimator as $\tilde{\boldsymbol{\beta}}_n$.

2. The adaptive group Lasso estimator is obtained by searching $\boldsymbol{\beta}_n$ to minimize

$$L_{n2}(\boldsymbol{\beta}_n, \lambda_{n2}) = \|\boldsymbol{y} - \mathbb{Z}_{\text{OCMT}} \boldsymbol{\beta}_n\|^2 + \lambda_{n2} \sum_{j=1}^{\hat{p}_{\text{OCMT}}} \frac{1}{\left\|\tilde{\boldsymbol{\beta}}_{nj}\right\|} \left\|\boldsymbol{\beta}_{nj}\right\|,$$

   where we use the convention that $0/0 = 0$. Denote the above estimator as $\hat{\boldsymbol{\beta}}_n$.

The post-selection estimation proceeds as follows. Denote the selected regressor from the above procedure as $\boldsymbol{Z}_{\text{AGLASSO}}$, and similarly denote its B-spline basis and design matrix as $P^{m_n}(\boldsymbol{Z}_{\text{AGLASSO}})$ and $\mathbb{Z}_{\text{AGLASSO}}$, respectively. The post selection estimator is the OLS estimator of regressing $\boldsymbol{y}$ on $\mathbb{Z}_{\text{AGLASSO}}$, which is

$$\hat{\boldsymbol{\beta}}_{\text{post}} = \left(\mathbb{Z}'_{\text{AGLASSO}} \mathbb{Z}_{\text{AGLASSO}}\right)^{-1} \mathbb{Z}'_{\text{AGLASSO}} \boldsymbol{y}.$$

The final fitted model is

$$P^{m_n}(\boldsymbol{Z}_{\text{AGLASSO}})' \hat{\boldsymbol{\beta}}_{\text{post}}.$$

Theorem 3.2 *Suppose that Assumptions 1, 2', 3, 4, 5', 6, 7, 8, 9', 10, and 11 hold. Further,* $\lambda_{n1} \geq C\sqrt{n \log(p^{**} m_n)}$ *for a sufficient large* C, $\lambda_{n1} \ll \sqrt{n/m_n}$, *and* $m_n^{1/2} \log(p^{**} m_n) \ll \lambda_{n2} \ll n m_n^{-1/4}$. *Then*

(i) All signal variables are kept and all pseudo signals or noise variables are eliminated w.p.a.1;

(ii) The post OCMT estimation error satisfies

$$P^{m_n} \left(\mathbf{Z}_{AGLASSO} \right)' \hat{\boldsymbol{\beta}}_{post} - \sum_{j=1}^{p^*} f_j^* \left(X_j \right) = O_P \left(\left(m_n/n \right)^{1/2} \right).$$

The above theorem is almost the same as that in Huang, Horowitz and Wei (2010) including the requirements on λ_{n1} and λ_{n2}, with the exception that the procedure starts with the covariates post OCMT. Consequently, we only need to show that the adaptive group Lasso procedure is still valid in the post-OCMT situation, and the rest follows immediately from Huang, Horowitz and Wei (2010). In particular, under Assumption 7, the biases of the post OCMT estimators of the nonparametric additive components are asymptotically negligible so that their mean square errors (MSEs) are dominated by their asymptotic variances that are of order $O\left(m_n/n\right)$, which explains the result in Theorem 3.2(ii).

Our procedure enjoys the same theoretical property as the adaptive group Lasso. After the OCMT, the dimension of candidate variables is reduced dramatically. Thus the additional computation burden applying the post-OCMT adaptive group Lasso can be almost ignored. We note that the main advantage of our procedure is fast and reliable computation, which delivers better small sample performance as shown from our simulation studies. An implication of the above theorem is that the post-OCMT adaptive group procedure improves over the post-selection estimation results in CKP (Theorem 2 in CKP) because pseudo-signals are now eliminated with very high probability.

3.5 Diverging p^*

Allowing a diverging p^* (number of true signals) is possible. The only additional technical condition apart from the restriction on the speed of p^* is that $\sum_{j=1}^{p^*} f_j^* \left(X_j \right)$ is uniformly bounded. Note that this condition naturally holds for a fixed p^* due to the boundedness of f_j^*. The main reason for the requirement of this condition is technical: the uniform boundedness of $\sum_{j=1}^{p^*} f_j^* \left(X_j \right)$ ensures that U_l defined in equation (2.5) also satisfies the exponential decayed tail condition,[3] and the tail property is necessary to apply the main inequalities to obtain the probability bounds. The uniform boundedness of $\sum_{j=1}^{p^*} f_j^* \left(X_j \right)$ was also imposed in Fan, Feng and Song (2011). Propositions 2.1 and 3.1 are on individual X_l, but we do need Assumption 2" on p^* so that Proposition 3.1 holds. It is to ensure that we can have precise estimation on a diverging design matrix.

Assumption 2" $p^* \lesssim n^{B_{p^*}}$ for some $B_{p^*} \geq 0$. $p^{**} \lesssim n^{B_{p^{**}}}$ and $p_n \propto n^{B_p}$ for some $B_p > B_{p^{**}}, B_{p^*} \geq 0$. Further, $B_{p^*} + B_{p^{**}} < \left(1 - 3B_m\right)/2$.

We present the main results in the following theorem.

Theorem 3.3 *Suppose that Assumptions 1, 2", 3, 4, 5', 6, 7, 8, 9', 10, and 11 hold. In addition, assume that $\sum_{j=1}^{p^*} f_j^* \left(X_j \right)$ is uniformly bounded. Then, the results in Theorem 3.1 continue to hold.*

[3] For details, see the proof of Lemma A.5.

In the next section, we investigate the small sample performance of our procedure by means of Monte Carlo experiment.

4 Monte Carlo Simulations

To investigate the finite-sample performance of our procedure, we conduct Monte Carlo experiments in this section.

4.1 Simulation Design

Following Huang, Horowitz and Wei (2010), we consider the following data generating processes (DGPs). In what follows, we assume

$$f_1(x) = x;\ f_2(x) = (2x-1)^2;\ f_3(x) = \frac{\sin(2\pi x)}{2-\sin(2\pi x)};\ \text{and}$$
$$f_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3.$$

For the errors, we assume $\varepsilon \sim$ i.i.d. $N(0,1)$ for DGPs 1–6 and 9–10. In DGPs 7–8, we check the impact of heteroskedastic errors on our methods. Specifically, we add a heteroskedastic error to the simplest and the most complicated designs in DGPs 1–6 to form DGPs 7 and 8, respectively. In DGPs 9–10, we consider the case with additive components of binary variables that mimic the application in Section 5.

DGP 1: Four independent signals only. Y is generated as follows:

$$Y = 2.55f_1(X_1) + 2.57f_2(X_2) + 1.68f_3(X_3) + f_4(X_4) + \varepsilon. \tag{4.1}$$

Note the coefficients before f_i's are set to make each signal have the same strength in terms of variance for independent uniform $X_1, \ldots, X_4$. The covariates are generated as follows:

$$X_j = W_j \text{ for } j = 1,\ldots,4, \text{ and } X_j = \frac{W_j + U_1}{2} \text{for } j \geq 5,$$

where W_j, $j = 1, \ldots, p_n$, and U_1 are all independent draws from $U(0,1)$. Thus, $p^* = 4$ and $p^{**} = 0$ for DGP 1. Define the Signal-to-noise ratio to be $r_{sn} = \frac{\mathtt{sd}(f)}{\mathtt{sd}(\varepsilon)}$, and $r_{sn} = 1.5$ for DGP 1.

DGP 2: Four independent signals and two pseudo-signals. Y is generated from equation (4.1). The covariates are generated as follows:

$$X_j = W_j \text{ for } j = 1,\ldots,4,\ X_5 = \frac{4X_1 + U_1}{5},\ X_6 = \frac{4X_2 + U_2}{5},\ \text{and}$$
$$X_j = \frac{W_{j-2} + U_3}{2} \text{ for } j \geq 7,$$

where W_j, $j = 1, \ldots, p_n - 2$, U_1, and U_2 are all independent draws from $U(0,1)$. Thus, $p^* = 4$ and $p^{**} = 2$ for DGP 2.

DGP 3: Four signals, and one hidden signal. Y is generated from

$$Y = 2.55 f_1(X_1) + 2.57 f_2(X_2) + 1.68 f_3(X_3) + f_4(X_4) + f_5(X_5) + \varepsilon, \tag{4.2}$$

where $f_5(X_5) = -\mathbb{E}\left[2.55 f_1(X_1) + 2.57 f_2(X_2) + 1.68 f_3(X_3) + f_4(X_4) \,|X_5\right]$. The covariates are generated as follows :

$$X_j = W_j \text{ for } j = 1, 2,\ X_j = \frac{W_j + U_1}{2} \text{ for } j = 3, 4,\ X_5 = U_1, \text{ and}$$
$$X_j = \frac{W_{j-1} + U_2}{2} \text{ for } j \geq 6,$$

where W_j, $j = 1, \ldots, p_n - 1$, U_1, and U_2 are independent draws from $U(0,1)$. Then, $p^* = 5$ and $p^{**} = 0$ for DGP 3, and the fifth signal is hidden by our definition.[4]

DGP 4: Four signals, two pseudo-signals, and one hidden signal. Y is generated from equation (4.2). The covariates are generated as follows :

$$X_j = W_j \text{ for } j = 1, 2,\ X_j = \frac{W_j + U_1}{2} \text{ for } j = 3, 4,\ X_5 = U_1,\ X_6 = \frac{4X_1 + U_2}{5},$$
$$X_7 = \frac{4X_2 + U_3}{5}, \text{ and } X_j = \frac{W_{j-3} + U_3}{2} \text{ for } j \geq 8,$$

where W_j, $j = 1, \ldots, p_n - 3$, U_1, U_2, and U_3 are independent draws from $U(0,1)$. Then, $p^* = 5$ and $p^{**} = 2$ for DGP 4, and the fifth signal is a hidden signal.

DGP 5: Four correlated signals. Y is generated from equation (4.1). The covariates are generated as follows:

$$X_j = \frac{W_j + U_1}{2} \text{ for } j = 1, \ldots, 4, \text{ and } X_j = \frac{W_j + U_2}{2} \text{ for } j \geq 5,$$

where W_j, $j = 1, \ldots, p_n$, U_1, and U_2 are independent draws from $U(0,1)$. Thus, four signals are correlated with each other, and $p^* = 4$ and $p^{**} = 0$ for DGP 5.

DGP 6: Four signals, many pseudo-signals, and one hidden signal. Y is generated from equation (4.2). The covariates are generated as follows :

$$X_j = W_j \text{ for } j = 1, 2,\ X_j = \frac{W_j + U_1}{2} \text{ for } j = 3, 4,\ X_5 = U_1,$$

[4] By the distribution of covariates,

$$f_5(x) \approx 0.97\pi - 1.2\cos(\pi x) + \sin(\pi x) + 0.6861\pi \arctan[2(\tan(\pi x/2) - 1)/\sqrt{3}]$$
$$- 0.6861\pi \arctan[2(\tan(\pi x/2 + \pi/2) - 1)/\sqrt{3}] + 0.2778\cos^3(\pi x) - 0.2222\sin^3(\pi x).$$

$$X_j = \frac{4X_1 + (j-5)\,W_{j-1}}{j-1} \text{ for } j = 6, 10, 14, 18, \ldots,$$
$$X_j = \frac{4X_2 + (j-5)\,W_{j-1}}{j-1} \text{ for } j = 7, 11, 15, 19, \ldots,$$
$$X_j = \frac{4X_3 + (j-5)\,W_{j-1}}{j-1} \text{ for } j = 8, 12, 16, 20, \ldots, \text{ and}$$
$$X_j = \frac{4X_4 + (j-5)\,W_{j-1}}{j-1} \text{ for } j = 9, 13, 17, 21, \ldots$$

where W_j, $j = 1, \ldots, p_n - 1$, and U_1 are independent draws from $U(0,1)$. Then, $p^* = 5$ with one hidden signal for DGP 6.

DGP 7: Four independent signals with heteroskedastic errors. Y is generated from equation (4.1) with the same covariates as in DGP 1. We assume that conditioning on X, ε is normal with mean 0 and variance $0.436[1 + (X_1 + X_2 + X_3 + X_4)/4]^2$, and the unconditional variance of ε is approximately 1.

DGP 8: Four signals, many pseudo-signals, and one hidden signal with heteroskedastic errors. Y is generated from equation (4.2) with the same covariates as in DGP 6. We assume that conditioning on X, ε is normal with mean 0 and variance $0.436[1 + (X_1 + X_2 + X_3 + X_4)/4]^2$.

DGP 9: Four independent signals with some binary variables. To mimic the application, we consider the situation with some binary covariates. Note that any function of a binary covariate can at most take two values. Without loss of generality, we focus on the case in which those binary covariates enter the model linearly. It is easy to see that our theoretical results continue to hold in the presence of some linear additive components with the main difference that they do not exhibit any approximation bias.[5] We set the threshold as $\varsigma_n = C\,[\log p_n]^{1.1}$. Y is generated from

$$Y = 2.57 f_2(X_1) + 1.68 f_3(X_2) + 1.47 X_3 + 1.47 X_4 + \varepsilon,$$

where $X_j = W_j$, $j = 1, 2$, $X_j = V_{j-2}$, $j = 3, 4$, W_1 and W_2 are independent $U(0,1)$, and V_1 and V_2 are independent Bernoulli random variables with equal chances of taking value 0 or 1. The remaining covariates are generated as follows:

$$X_j = \frac{W_{j-2} + U_1}{2} \text{ for } j = 5, 6, \ldots, \frac{p_n}{2}, \text{ and } X_j = V_{j - \frac{p_n}{2} + 2} \text{ for } j = \frac{p_n}{2} + 1, \ldots, p_n,$$

where W_j, $j = 5, 6, \ldots, \frac{p_n}{2} - 2$, and U_1, are independent $U(0,1)$, V_j, $j = 3, \ldots, \frac{p_n}{2} + 2$ are distributed the same as V_1. All Ws, Vs, and U are independent of each other.

DGP 10: Four signals with one hidden signal in the presence of some binary variables. For this DGP, Y is generated from

$$Y = 2.57 f_2(X_1) + 1.5 X_2 + 1.5 X_3 - X_4 + \varepsilon,$$

[5] In this case, the test statistics for the additive linear terms are the squares of t-statistics, and the inequality continues to hold by setting, for example, $\varsigma_n \propto [\log p_n]^{1.1}$ for the case when p_n is a polynomial of n.

where $X_1 = W_1$ is a $U(0,1)$, $X_j = V_{j-1}$, $j = 2,3,4$, are Bernoulli random variables with equal chances of taking 0 or 1, and Corr(V_1, V_3) =Corr(V_1, V_2) =Corr$(V_2, V_3) = 1/3$. X_j, $j = 5, 6, ..., p_n$ are the same as those in DGP 9.[6] Some simple calculation implies that X_4 is a hidden signal.

4.2 Tuning Parameters

The key tuning parameter in this study is the threshold ς_n. The CKP's Bonferroni correction strategy does not perform consistently well for our case. We do the following instead. We set

$$\varsigma_n = C m_n \left[(\log p_n)^{1.1} + (\log m_n)^{1.1} \right] \text{ and } \varsigma_n = C (\log p_n)^{1.1}$$

for continuous variables and binary variables, respectively. This satisfies Assumption 8 when, in addition, Assumptions 2 and 7 hold (both m_n and p_n grow at the polynomial rate of n). In the small samples, we set this ς_n to have control over both m_n and p_n. CKP suggest using a larger threshold, ς_n^*, for subsequent stages, to improve the finite sample performance. We follow their lead and set a ς_n^* larger than ς_n for subsequent stages. That is, we replace ς_n with ς_n^* for $\widehat{\mathcal{J}}_{l,(2)}, \widehat{\mathcal{J}}_{l,(3)}, ..., \widehat{\mathcal{J}}_{l,(k)}$ in Section 3.3. The reason is that the chance of including a noise variable increases quickly for subsequent stages, and we need a larger ς_n^* for the OCMT to conclude more easily. We set $\varsigma_n^* = 3\varsigma_n$ for continuous variables. Note that CKP take the threshold for later stages to be twice the threshold for the first stage, and their test is based on t-statistics. Since ours is based on chi-squared statistics, equivalently, we should set ς_n^* to be $4\varsigma_n$. However, some small-scale experiments suggest that setting $\varsigma_n^* = 3\varsigma_n$ can yield better small sample performance, probably because our model is nonparametric and our test statistic varies more than the parametric counterpart. Note that this change does not affect our theoretical results because ς_n^* is proportional to ς_n. For binary additive components, we continue to follow CKP and set $\varsigma_n^* = 4\varsigma_n$ because they enter the model linearly.

Another important tuning parameter is m_n, which is critical for the sieve estimation. The optimal choice of m_n has been studied extensively in the literature. Popular ways to choose the value of m_n include cross validation, Akaike information criterion (AIC), and BIC. We refer the readers to Chen (2007) and Hansen (2014) for a review on this important issue. For m_n, we simply set $m_n = \lfloor n^{1/4} \rfloor + 1$, where $\lfloor \cdot \rfloor$ is the floor operator. This m_n satisfies Assumption 7, if $d > 3/2$. Other choices of sieve terms such as $m_n = \lfloor n^{1/4} \rfloor + 2$ are also considered in the simulations, and they yield similar results.

For the C in $\varsigma_n = C m_n \left[(\log p_n)^{1.1} + (\log m_n)^{1.1} \right]$ or $\varsigma_n = C (\log p_n)^{1.1}$, we test C in the range of 0.5 to 2.5, specifically, $0.5, 0.6, ..., 2.5$. We determine the value of C by minimizing the following BIC:

$$\text{BIC}(C) = n \log [\text{RSS}(C)/n] + (\text{number of selected variables}) \cdot \log(n), \quad (4.3)$$

[6] An example of V_1, V_2, and V_3 is that $V_j = \mathbf{1}(\tilde{U}_j + \tilde{U}_4 > 1)$ for $j = 1, 2$, and 3, where $\tilde{U}_1, ..., \tilde{U}_4$ are independent $U(0,1)$.

where RSS(C) denotes the residual sum of squares from the post-OCMT ordinary least squares regression of Y on $P^{m_n}\left[\boldsymbol{Z}_{(\hat{k}_s)}\right]$, where $\boldsymbol{Z}_{(\hat{k}_s)}$ are the variables selected by the OCMT with C in use. The tuning parameters for the adaptive group Lasso are selected as in Section 4.3. Another popular way to choose C is cross validation. As seen from the results, the BIC works well for our procedure. In light of this, we will not pursue the procedure with cross validation.

For an easy reference, we present the implementation details in Appendix A.1.

4.3 Estimators Compared

For comparison, we consider the adaptive group Lasso by Huang, Horowitz and Wei (2010), which is designed for component selection in the nonparametric additive model. It is a two-step approach—the first step is the usual group Lasso and the second is the adaptive group Lasso with initial estimates from the first step. We select tuning parameters λ_{n1} in step 1 and λ_{n2} in step 2 (λ_{n1} and λ_{n2} are in the notations of Huang, Horowitz and Wei (2010)) by BIC as well. Specifically, we set $\lambda_{n1} = \lambda_j$ with

$$\lambda_j = \exp\left\{\log\left(\lambda_{\max}\right) + \left[\log\left(\lambda_{\min}\right) - \log\left(\lambda_{\max}\right)\right]\frac{j}{30}\right\}$$

for $j = 0, 1, ..., 30$, where $\lambda_{\min} = \max\left\{0.05, 10^{-5}\left\|\boldsymbol{y}\right\|\right\}$ and $\lambda_{\max} = 0.5\left\|\boldsymbol{y}\right\|$. We calculate BIC$_j$ based on the estimation using $\lambda_{n1} = \lambda_j$, and we select the $\hat{\lambda}_{n1}$ that minimizes the BIC$_j$ among $j = 0, 1, ..., 30$. We set the estimates in the first step as the estimates using $\hat{\lambda}_{n1}$. For the second step, we set $\lambda_{n2} = \lambda_j$ for $j = 0, 1, ..., 30$ with the initial estimates as the estimates from the first step using $\hat{\lambda}_{n1}$. We again calculate BIC$_j$ based on the estimation using $\lambda_{n2} = \lambda_j$, and we select the $\hat{\lambda}_{n2}$ that minimizes the BIC$_j$ among $j = 0, 1, ..., 30$. The final estimates are the adaptive group Lasso estimates using $\hat{\lambda}_{n2}$. Note that the variables not selected in the first step are not included in the second step, because the penalty for those variables is infinity in the second step. We find the solutions in both steps through the block coordinate descent algorithm (i.e., the "shooting" algorithm). For the algorithm details, see Wu and Lange (2008).[7]

In addition, we compare our method with the random forest regression and bagging, both of which are commonly-used machine learning methods. Since there is no variable selection criterion in random forest, we focus on the comparison of out-sample forecasting performance.

4.4 Estimation Results

We consider the combinations of $n = 200$ or 400 and $p_n = 100$, 200, or 1,000 for each DGP. All results are based on 1,000 replications. We report the results for five different methods. The first and second

[7] One implementation in MATLAB can be found at: https://publish.illinois.edu/xiaohuichen/code/group-lasso-shooting/.

methods are our post-OCMT procedure (Steps 1–6 in Appendix A.1, denoted as "POST–OCMT") and OCMT procedure (Steps 1– 5 in Appendix A.1, denoted as "OCMT"), respectively. The third method is the adaptive group Lasso (denoted as "AGLASSO"), and the last two methods are bagging (denoted as "BAGGING") and random forest (denoted as "RF"), respectively.

Following the literature, we report the mean number of variables selected (NV), the true positive rates (TPR), the false positive rates (FPR), the false discovery rates (FDR), the percentage of correct selection (CS) for the first three methods[8], and the out-of-sample root mean squared forecast errors (RMSFE) for all the five methods.[9] We report the average number of stages (STEP) for our OCMT procedure only.

To save space, we report the results in the main body of this paper for only DGPs 1 and 6 in Tables 3 and 4, respectively. These two designs correspond to the simplest and the most complicated designs in the simulation, respectively. Results for DGPs 2–5 and 7–10 are provided in Tables S1 to S8 in Appendix S2. To showcase the advantage of the post-OCMT procedure compared to the OCMT procedure only, we also consider a linear DGP (DGP 11) in the online Appendix S2 with results reported in Table S9.

We can make several observations. First, POST–OCMT performs the best among the three methods in most cases. POST–OCMT outperforms OCMT due to its ability to eliminate pseudo-signals or noise variables. POST–OCMT even outperforms OCMT slightly for DGPs without any pseudo-signals, especially for $n = 200$. This result confirms the necessity of conducting the additional post-OCMT step. Moreover, the additional computation time for POST–OCMT compared with OCMT can be almost ignored because the number of candidate variables after OCMT is very small. Second, AGLASSO performs almost the same as our procedures for DGP 2. Note that DGP 2 contains four independent signals, two pseudo-signals, and no hidden signals. Then, this result is not surprising, because Lasso performs very well for independent signals, and the biggest challenge for our procedure is the possible presence of pseudo-signals. When the signals are correlated in DGP 4, AGLASSO performs less well and is outperformed by POST–OCMT. Third, the performance of all methods improves as n increases from 200 to 400. Fourth, OCMT is very successful at picking up hidden signals, especially for $n = 400$; see, for example, the results for DGPs 3, 4, 6, and 8. Fifth, our methods perform well in the presence of heteroskedastic errors for DGP 7 and 8. Sixth, the CS of POST–OCMT performs well even for the complicated DGPs 6 and 8 at $n = 400$, whereas AGLASSO performs poorly in terms of CS for these two DGPs. Seventh, the number of stages for

[8] That is, we precisely uncover the true model in (2.3) using all the true signals, but not any pseudo-signals or noise variables.

[9] The forecasts are based on the post-selection estimates from each method. In each replication, we additionally independently generate 200 observations. Then, we calculate the root mean square errors (RMSE) of the difference between the forecasts and Y for the new observations. RMSFE is the average of those RMSE for 1,000 replications.

Table 3: DGP 1

Panel 1: $n = 200$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	4.0080	0.9998	0.0001	0.0015	0.9910	-	1.0718
OCMT	4.0180	0.9998	0.0002	0.0031	0.9820	1.0260	1.0886
AGLASSO	4.0410	0.9990	0.0005	0.0075	0.9550	-	1.0753
BAGGING	-	-	-	-	-	-	1.4403
RF	-	-	-	-	-	-	1.4846
Panel 2: $n = 200$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	4.0150	0.9998	0.0001	0.0027	0.9830	-	1.0740
OCMT	4.0200	0.9998	0.0001	0.0035	0.9790	1.0390	1.0939
AGLASSO	4.0510	0.9950	0.0004	0.0114	0.9330	-	1.0808
BAGGING	-	-	-	-	-	-	1.4690
RF	-	-	-	-	-	-	1.5160
Panel 3: $n = 200$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	4.0010	0.9952	0.0000	0.0034	0.9640	-	1.0773
OCMT	4.0200	0.9952	0.0000	0.0062	0.9510	1.0880	1.1173
AGLASSO	3.9370	0.9620	0.0001	0.0145	0.8770	-	1.0993
BAGGING	-	-	-	-	-	-	1.5281
RF	-	-	-	-	-	-	1.5847
Panel 4: $n = 400$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	4.0010	1.0000	0.0000	0.0002	0.9990	-	1.0470
OCMT	4.0010	1.0000	0.0000	0.0002	0.9990	1.0000	1.0470
AGLASSO	4.0080	0.9990	0.0001	0.0020	0.9870	-	1.0531
BAGGING	-	-	-	-	-	-	1.3234
RF	-	-	-	-	-	-	1.3648
Panel 5: $n = 400$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	4.0010	1.0000	0.0000	0.0002	0.9990	-	1.0477
OCMT	4.0010	1.0000	0.0000	0.0002	0.9990	1.0000	1.0477
AGLASSO	4.0210	1.0000	0.0001	0.0035	0.9790	-	1.0520
BAGGING	-	-	-	-	-	-	1.3536
RF	-	-	-	-	-	-	1.3984
Panel 6: $n = 400$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	4.0040	1.0000	0.0000	0.0007	0.9960	-	1.0519
OCMT	4.0040	1.0000	0.0000	0.0007	0.9960	1.0000	1.0519
AGALSSO	4.0313	1.0000	0.0000	0.0052	0.9688	-	1.0524
BAGGING	-	-	-	-	-	-	1.4120
RF	-	-	-	-	-	-	1.4574

Table 4: DGP 6

Panel 1: $n = 200$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	3.6300	0.6628	0.0033	0.0650	0.1770	-	1.2296
OCMT	4.8450	0.6834	0.0149	0.2163	0.0000	1.2650	1.4003
AGLASSO	2.9840	0.5522	0.0023	0.0432	0.0040	-	1.3670
BAGGING	-	-	-	-	-	-	1.4522
RF	-	-	-	-	-	-	1.4645
Panel 2: $n = 200$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	3.5600	0.6438	0.0017	0.0682	0.1290	-	1.2396
OCMT	4.6700	0.6642	0.0069	0.2107	0.0000	1.2220	1.3861
AGLASSO	2.6340	0.4866	0.0010	0.0417	0.0000	-	1.3695
BAGGING	-	-	-	-	-	-	1.4855
RF	-	-	-	-	-	-	1.4969
Panel 3: $n = 200$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	3.2550	0.5892	0.0003	0.0669	0.0440	-	1.2674
OCMT	4.2310	0.6088	0.0012	0.2040	0.0000	1.0880	1.3311
AGLASSO	1.9780	0.3678	0.0001	0.0306	0.0000	-	1.4304
BAGGING	-	-	-	-	-	-	1.5395
RF	-	-	-	-	-	-	1.5539
Panel 4: $n = 400$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	4.8840	0.9448	0.0017	0.0246	0.6960	-	1.0659
OCMT	7.4660	0.9534	0.0281	0.3056	0.0000	1.8380	1.4894
AGLASSO	4.3480	0.8276	0.0022	0.0331	0.1060	-	1.1294
BAGGING	-	-	-	-	-	-	1.3502
RF	-	-	-	-	-	-	1.3767
Panel 5: $n = 400$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	4.8870	0.9488	0.0007	0.0213	0.7080	-	1.0723
OCMT	7.4120	0.9542	0.0135	0.3019	0.0000	1.8360	1.5018
AGLASSO	4.2240	0.8122	0.0008	0.0270	0.0630	-	1.1410
BAGGING	-	-	-	-	-	-	1.3848
RF	-	-	-	-	-	-	1.4118
Panel 6: $n = 400$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST-OCMT	4.8380	0.9356	0.0002	0.0242	0.6740	-	1.0784
OCMT	7.4140	0.9398	0.0027	0.3082	0.0000	1.8030	1.4863
AGLASSO	4.1340	0.7916	0.0002	0.0299	0.0110	-	1.1509
BAGGING	-	-	-	-	-	-	1.4477
RF	-	-	-	-	-	-	1.4689

OCMT basically confirms our theoretical findings. For example, for DGP 1, the mean number of stages is slightly more than 1 for $n = 200$, and is 1 for $n = 400$. In the presence of hidden signals, the mean number of stages is approximately 2 for $n = 400$ for DGPs 3, 4, 6, 8, and 10. Note the mean number of stages is approximately 2 for $n = 400$ for DGP 5 with correlated signals and no hidden signals. The reason is that the correlation makes the net effect of X_1 on Y very small and X_1 almost behaves like a "hidden signal". This result also confirms the necessity of using multiple stages instead of a single-stage procedure. Eighth, our method also works well for DGPs 9 and 10 that mimic the application. An additional remark is that our procedure can be implemented fast and is much faster than AGLASSO. For example, when $p = 1,000$, our procedure took less than half a minute for one replication on average, whereas the AGLASSO took hours for one replication. Finally, the first three procedures have smaller RMSFE than Bagging and RF in almost all scenarios and thus have better out-of-sample forecasting performance.[10] This is because they explicitly utilize the information (in the form of an additive function) underlying the DGPs while BAGGING and RF do not. In addition, POST–OCMT delivers the best out-of-sample forecasting performance among all procedures.

To summarize, our methods perform well in small samples, and we view it as a useful alternative to existing methods in the literature.

5 An Application

In this section, we apply our method to a dataset extracted from RUMiC.[11] The survey studied immigrants or workers moving from the rural areas of China to its big cities. The survey asked interviewees (immigrants or workers) a wide range of questions. For the detailed design of the survey and other information, including on the construction of each variable, see the survey website and Gong et al. (2008). Currently, the 2008 wave data are publicly available.

Economic reforms since the late 1970s have brought significant changes to China's economy. The government began relaxing its policy on population mobility in the early 1980s. Gradually, peasants were allowed to leave villages and work in big cities to earn higher incomes. Most migrant workers may leave their spouses, children, or parents behind in their hometowns who may need their financial

[10] Note that Bagging has slightly better out-of-sample forecasting performance than RF. This is reasonable given that our DGPs have a finite number of signal variables. In RF, many decision splits do not improve predictive accuracy because they rely solely on noise variables to generate the trees.

[11] RUMiC consists of three parts: the Urban Household Survey, the Rural Household Survey, and the Migrant Household Survey. A group of researchers at the Australian National University, the University of Queensland, and the Beijing Normal University initiated this survey. The Institute for the Study of Labor (IZA) supported it and provides the Scientific Use Files. RUMiC had financial support from the Australian Research Council, the Australian Agency for International Development, the Ford Foundation, IZA, and the Chinese Foundation of Social Sciences. More information on the survey can be found at https://datasets.iza.org/dataset/58/longitudinal-survey-on-rural-urban-migration-in-china.

support. This situation results in monetary transfers, that is, remittances, from migrant workers to their family. In the context of migration, family, and economic development, remittances are not only an income source for recipients, but also reflect intrafamilial relationships. Remittances clearly represent a dimension of family ties and demonstrate high degrees of interaction between migrants and families at home. In addition, remittances from the rural migrant workers also contribute significantly to China's agricultural productivity (c.f., Rozelle, Taylor and deBrauw (1999)). For these reasons, it has long been of interest to model remittances to families or relatives in the hometown; see Li (2001) and Cai (2003), among others.

In this application, we take remittance as the dependent variable (Y); we focus on the dataset from Guangdong Province (Guangzhou, Dongguan, and Shenzhen cities) in the 2008 survey wave, and keep 78 covariates from the dataset.[12] After dropping observations with missing information, the number of observations is 456. We provide the definitions of the dependent variable and covariates, and the associated summary statistics, in Tables S10 and S11, respectively. We report the original labels of all covariates in the survey in the first column of Table S10 for reference. Among the 78 covariates, there are some continuous variables with most observations as 0 (Panel C in Table S10), and some discrete variables with very limited support (Panel D in Table S10). For those variables, the design matrix of the sieves generated are either singular or close to singular. For this reason, we add those variables linearly into the model and treat them the same as dummy variables (Panel E in Table S10) for modeling. Consequently, the way we fit the dataset resembles the approach we used for DGPs 9 and 10 in the simulation. We explain the reason our theoretical results continue to hold in this situation in the simulation section (DGP 9). We take natural logarithms for the Y and continuous X variables to offset the effect of outliers; otherwise, the forecast can easily take some extreme values.

We randomly select 400 observations as the training sample, and the remaining 56 observations as the test sample. The number of sieve terms is set as $m_n = \lfloor 400^{1/4} \rfloor + 1 = 5$ for the continuous variables in Panel B of Table S10). We set $\varsigma_n = Cm_n \left[(\log p_n)^{1.1} + (\log m_n)^{1.1} \right]$ and $\varsigma_n^* = 3\varsigma_n$ for continuous variables, and $\varsigma_n = C (\log p_n)^{1.1}$ and $\varsigma_n^* = 4\varsigma_n$ for terms entering the model linearly (see Section 4.2 for the reason). We set C in the range of 0.5 to 2.5, specifically, $0.5, 0.6, ..., 2.5$, and choose C to minimize the BIC for the model selection. The competing methods are the group Lasso (labelled as GLASSO) and the adaptive group Lasso (AGLASSO). The tuning parameters for GLASSO and AGLASSO are selected as in Section 4.3. We evaluate the performance of all methods based on the RMSFE of the test dataset using the fitted models from different methods. We independently repeat the above procedure 100 times.

We report the results in terms of the out-of-sample RMSFE in Tables 5 and 6, with AGLASSO

[12] We keep covariates with relatively fewer missing observations.

Table 5: Performance in Terms of RMSFE (100 Cases), benchmark: AGLASSO

OCMT			POST-OCMT			GLASSO		
Better	Same	Worse	Better	Same	Worse	Better	Same	Worse
77	17	6	76	21	3	5	88	7

Table 6: Average RMSFE ratio, benchmark: AGLASSO

One Stage	OCMT	POST-OCMT	GLASSO	AGLASSO
0.873	0.817	0.814	1.005	1

as the benchmark. Table 5 shows that OCMT and POST-OCMT outperform AGLASSO in the majority of cases. Of course, our methods do not outperform AGLASSO all the time. To highlight the necessity of the multiple stages, we also report the results of the one-stage procedure (we selected the tuning parameter also by minimizing the BIC) in Table 6. The OCMT stops at the second stage for all 100 cases. We normalize the average RMSFE of AGLASSO as 1. Notably, the average RMSFE of our methods are lower than that of AGLASSO. The OCMT also improves the RMSFE over the one-stage procedure, possibly owing to some hidden signals uncovered by our multiple-stage procedure. We report the frequencies of variables (out of 100 cases) selected by all methods in Table S12. It appears that G102 (monthly income) along with G133 (gifts to others, including parents) and G137 (education cost for left-behind children) contribute most to the model, as shown by all methods in general. We note that AGLASSO tends to select more variables than our methods, on average, which was also CKP's finding in their application. The "over-fitting" is the main cause of the relative inferior performance of AGLASSO.

6 Conclusion

In this paper, we examine the one-covariate-at-a-time multiple testing approach to model selection in additive models. The properties of the TPR, FPR, and FDR of our approach are established based on some asymptotic probability bounds of Type-I and II errors. The simulation experiments and one application on the RUMiC dataset showcase excellent small-sample properties of our methods. Just as stated by CKP for linear models, we view our approach as a useful alternative to the model selection methods for additive models in the literature.

Appendix

In Appendix A.1, we present a detailed procedure for practice. We then provide technical details for the one-stage and multiple-stage procedures in Appendices A.2 and A.3, respectively. Additionally, we include technical lemmas used in the proofs of the main results in Appendix A.4. Finally, the main results of the paper are proven in Appendix B. The proofs of the technical lemmas and some additional simulation and application results are available in the online supplement.

A The Procedure, Some Technical Details and Technical Lemmas

A.1 The Procedure

For easy reference, we summarize how to implement our procedure in this section. We refer readers to the justification of our selection of tuning parameters in Section 4.2. We begin with an introduction to B-splines. Then we move to the details of the implementation.

A.1.1 A Gentle Introduction to B-splines

Here, we provide a brief introduction to B-splines. For more details on regression splines, see Racine (2022). B-splines are defined by their "order" m and the number of interior "knots" N. The degree (J) of the B-spline polynomial is given by the spline order m minus one, i.e., $J = m - 1$.

Let $t_0 \leq t_1 \leq \cdots \leq t_N \leq t_{N+1}$ be the knot set, where t_0 and t_{N+1} are two "endpoints" knots, and $t_1, \ldots, t_N$ are N interior knots. The splines with equidistant knots are called "*uniform*" splines; otherwise, they are said to be "*nonuniform*". Another popular choice of knots is the "quantile" knot sequence, where the empirical quantiles of the observable variable are used as interior knots. Quantile knots ensure an equal number of sample observations in each interval, while the intervals can have different lengths.

To construct the B-spline basis function recursively, define the augmented knot set by appending the lower and upper boundary knots t_0 and t_{N+1} $J\,(= m-1)$ times:

$$t_{-(m-1)} = \cdots = t_{-1} = t_0 \leq t_1 \leq \cdots \leq t_N \leq t_{N+1} = t_{N+2} = \cdots = t_{N+m}.$$

Reset the index for the augmented knot set such that the $N+2m$ augmented knots t_i are now indexed by $i = 0, ..., N+2m-1$. Then we can recursively define basis functions $B_{i,j}$, $j = 0, 1, ..., J$, as follows:

$$B_{i,0}(x) = \begin{cases} 1 \text{ if } x \in [t_i, t_{i+1}) \\ 0 \text{ otherwise,} \end{cases}$$

$$B_{i,j+1}(x) = c_{i,j+1}(x) B_{i,j}(x) + [1 - c_{i+1,j+1}(x)] B_{i+1,j}(x),$$

where $c_{i,j}(x) = \frac{x - t_i}{t_{i+j} - t_i}$ if $t_{i+j} \neq t_i$; 0 otherwise.

The above recurrence relation is called the *de Boor recurrence relation*; see de Boor (2001). For a fixed j, the functions $B_{i,j}$ are called the ith B-spline basis functions of degree j, and the total number of $B_{i,j}(x)$ functions is $N+j+1$, according to the recursive construction. Finally, the B-spline basis functions used in regression are $\{B_{i,m-1}(x)\}_{i=0}^{N+m}$, and the total number of basis functions is $N+m$. When $m=4$ and $N=3$, there are 7 cubic B-spline basis functions.

In our simulation and application, we choose cubic B-splines with $m=4$ and set the number of interior knots to be $N=m_n-4$, where $m_n=\lfloor n^{1/4}\rfloor+1$ is the total number of basis functions. Here $\lfloor\cdot\rfloor$ is the floor operator.

A.1.2 The Detailed Procedure

Preparation

1. Normalize the dependent variable to have sample mean 0.

2. For each covariate, say X_l, normalize the B-spline basis to mean 0:

$$\phi_{jl}\left(x\right)=\psi_j\left(x\right)-n^{-1}\sum_{i=1}^{n}\psi_j\left(x_{li}\right),$$

   for $j=1,2,...,m_n$.

3. Collect the B-spline basis for covariate X_l as $P^{m_n}\left(X_l\right)=\left[\phi_{1l}\left(X_l\right),\phi_{2l}\left(X_l\right),...,\phi_{m_nl}\left(X_l\right)\right]'$.

An Instruction on Regression

1. To run the regression of Y on X_l, we regress the demeaned dependent variable Y on $P^{m_n}\left(X_l\right)$. We construct $\hat{\mathcal{X}}_l$ as in equation (2.6).

2. To run the regression of Y on X_l with pre-selected variables $\boldsymbol{Z}$, we obtain the estimator by the partitioned regression of Y on $P^{m_n}\left(X_l\right)$ by partialling out $P^{m_n}\left(\boldsymbol{Z}\right)$. $\hat{\mathcal{X}}_{l,\boldsymbol{Z}}$ is constructed as in equation (3.2).

Tuning Parameters For a fixed n, we set $m_n=\lfloor n^{1/4}\rfloor+1$. We set

$$\varsigma_n=Cm_n\left[\left(\log p_n\right)^{1.1}+\left(\log m_n\right)^{1.1}\right]\text{ (used at Stage 1)}$$

and $\varsigma_n^*=3\varsigma_n$ (used at Stages 2 or later) for continuous regressors. We set $\varsigma_n=C\left(\log p_n\right)^{1.1}$ (used at Stage 1) and $\varsigma_n^*=4\varsigma_n$ (used at Stages 2 or later) for binary regressors. We experiment with C in the range of 0.5 to 2.5, specifically using values such as 0.5, 0.6, ..., 2.5. We choose the optimal C by the BIC.

The Procedure

1. Set $\varsigma_n = 0.5 m_n \left[(\log p_n)^{1.1} + (\log m_n)^{1.1}\right]$ and $\varsigma_n^* = 3\varsigma_n$ for continuous variables, $\varsigma_n = 0.5 (\log p_n)^{1.1}$ and $\varsigma_n^* = 4\varsigma_n$ for binary variables, and then conduct the OCMT procedure as described later in this section;

2. Denote the selected variables from the OCMT procedure by $\boldsymbol{Z}_{(\hat{k}_s)}$;

3. Re-estimate the model by regressing Y on $P^{m_n}\left[\boldsymbol{Z}_{(\hat{k}_s)}\right]$ (which is a post-selection estimation);

4. Compute the BIC (defined in equation (4.3)) from Step 3 and denote it as BIC_1;

5. Repeat Steps 1 to 4, replacing $C = 0.5$ in the definition of ς_n and ς_n^* by $0.6, 0.7, \ldots, 2.5$ and obtain the corresponding BIC, denoted as $\text{BIC}_2, \text{BIC}_3, \ldots, \text{BIC}_{21}$;

6. Find $\hat{\imath} = \arg\min_{1\le\iota\le 21} \text{BIC}_\iota$ and select the variables obtained using the $\hat{\imath}$-th values of ς_n and ς_n^*;

7. Conduct the adaptive group Lasso (e.g., Section 4.3) for those variables selected in Step 6.

The OCMT Procedure This is a brief summary of Section 3.3 but is detailed enough for implementation.

1. At Stage 1, we regress demeaned Y on $P^{m_n}(X_l)$ one by one for $l = 1, 2, ..., p_n$, and calculate the test statistic $\hat{\mathcal{X}}_l$ as in equation (2.6). We select X_l if $\widehat{\mathcal{J}}_{l,(1)} = \mathbf{1}\left(\hat{\mathcal{X}}_l > \varsigma_n\right) = 1$.

2. We collect all the variables selected in stage 1 into the vector $\boldsymbol{Z}_{(1)}$, and denote the index set of the selected variables as $S_{(1)}$. The active variables in stage 2 are denoted by $\Psi_{(2)} = \{1, 2, \ldots, p_n\} \setminus S_{(1)}$.

3. For stage $k \ge 2$, we regress Y on $P^{m_n}(X_l)$ with $P^{m_n}\left(\boldsymbol{Z}_{(k-1)}\right)$ as pre-selected variables one by one for $l \in \Psi_{(k)}$. We construct the test statistic $\hat{\mathcal{X}}_{l,\boldsymbol{Z}_{(k-1)}}$ as in equation (3.2). We select the variable X_l if $\widehat{\mathcal{J}}_{l,(k)} = \mathbf{1}\left(\hat{\mathcal{X}}_{l,\boldsymbol{Z}_{(k-1)}} > \varsigma_n^*\right) = 1$.

4. We stop the procedure at a stage in which no new variables are selected. This final stage is denoted as $\hat{k}_s$. $\boldsymbol{Z}_{(\hat{k}_s)}$ are the variables selected by the OCMT.

A.2 Some Technical Details for the One-Stage Procedure

The proof of Proposition 2.1 is tedious. We illustrate the main idea of the proof by providing some brief technical details below.

Define

$$f_{nl}(x) = P^{m_n}(x)' \boldsymbol{\beta}_l, \text{ and } U_l = Y - f_{nl}(X_l). \tag{A.1}$$

Then $E[P^{m_n}(X_l) U_l] = 0$ by the definition of $f_{nl}(x)$. By Stone (1985), it holds that

$$\left\{ E[f_{nl}(X_l) - f_l(X_l)]^2 \right\}^{1/2} \le C_1 m_n^{-d} \text{ for some } C_1 < \infty. \tag{A.2}$$

If the bias is asymptotically negligible, by the properties listed in Lemma A.3 below,

$$\|\boldsymbol{\beta}_l\| \propto m_n^{1/2} \left\{ E\left[f_{nl}(X_l)^2 \right] \right\}^{1/2} \propto m_n^{1/2} \left\{ E\left[f_l(X_l)^2 \right] \right\}^{1/2} = m_n^{1/2} \theta_l.$$

As a result $m_n^{-1/2} \|\boldsymbol{\beta}_l\|$ can be equivalently used to measure the strength of net impact of X_l on Y.

Let $u_{li} = y_i - f_{nl}(x_{li})$ and $\boldsymbol{u}_l = (u_{l1}, u_{l2}, \ldots, u_{ln})'$. Then the estimator for $\boldsymbol{\beta}_l$ can be rewritten as

$$\hat{\boldsymbol{\beta}}_l = \left(\mathbb{X}_l' \mathbb{X}_l \right)^{-1} \mathbb{X}_l' \boldsymbol{y} = \boldsymbol{\beta}_l + \left(\mathbb{X}_l' \mathbb{X}_l \right)^{-1} \mathbb{X}_l' \boldsymbol{u}_l,$$

where recall that $\mathbb{X}_l = (\mathbb{X}_{l1}, \mathbb{X}_{l2}, \ldots, \mathbb{X}_{ln})'$ is the $n \times m_n$ "design" matrix for X_l. With it, $\hat{\mathcal{X}}_l$ can be rewritten as

$$\hat{\mathcal{X}}_l = \hat{\boldsymbol{\beta}}_l' \left(\hat{\sigma}_l^{-2} \mathbb{X}_l' \mathbb{X}_l \right) \hat{\boldsymbol{\beta}}_l = \boldsymbol{\beta}_l' \left(\hat{\sigma}_l^{-2} \mathbb{X}_l' \mathbb{X}_l \right) \boldsymbol{\beta}_l + 2 \hat{\sigma}_l^{-2} \boldsymbol{\beta}_l' \mathbb{X}_l' \boldsymbol{u}_l + \boldsymbol{u}_l' \mathbb{X}_l \left(\hat{\sigma}_l^2 \mathbb{X}_l' \mathbb{X}_l \right)^{-1} \mathbb{X}_l' \boldsymbol{u}_l. \tag{A.3}$$

Our analysis is based on the decomposition in (A.3). The idea is to show that $\left\| n^{-1/2} \mathbb{X}_l' \boldsymbol{u}_l \right\|$, $\hat{\sigma}_l^2$, and $\left\| n^{-1} \mathbb{X}_l' \mathbb{X}_l \right\|$ are bounded by some rate or value with very high probability. Based on these results, the value of $\widehat{\mathcal{J}}_l$ is solely determined by the strength of $\|\boldsymbol{\beta}_l\|$ with very high probability.

A.3 Some Technical Details for the Multiple-Stage Procedure

The proof of Proposition 3.1 is rather tedious. However, the main idea of the proof is straightforward, and similar to the one for Proposition 2.1. To see it, we provide some technical discussion below.

To simplify notations, let $P^{m_n}\left(\boldsymbol{X}_1^{p^*} \right) \equiv \left[P^{m_n}(X_1)', \ldots, P^{m_n}(X_{p^*})' \right]'$, a $p^* m_n \times 1$ vector. Define the $n \times p^* m_n$ matrix

$$\mathbb{X}_1^{p^*} \equiv (\mathbb{X}_1, \ldots, \mathbb{X}_{p^*}),$$

which is the design matrix for all the p^* signal variables. To see the approximation of the model more clearly, let

$$\tilde{\boldsymbol{\beta}}_1^{p^*} = \left\{ E\left[P^{m_n}\left(\boldsymbol{X}_1^{p^*} \right) P^{m_n}\left(\boldsymbol{X}_1^{p^*} \right)' \right] \right\}^{-1} E\left[P^{m_n}\left(\boldsymbol{X}_1^{p^*} \right) Y \right],$$

which is the population coefficient for the regression of Y on $P^{m_n}\left(\boldsymbol{X}_1^{p^*} \right)$. Write $\tilde{\boldsymbol{\beta}}_1^{p^*} = \left(\tilde{\boldsymbol{\beta}}_1', \ldots, \tilde{\boldsymbol{\beta}}_{p^*}' \right)'$, where $\tilde{\boldsymbol{\beta}}_j$ corresponds to the coefficient of $P^{m_n}(X_j)$ in the regression. Let $f_{nj}^*(x) = P^{m_n}(x)' \tilde{\boldsymbol{\beta}}_j$. Under Assumption 5', Stone (1985) implies

$$\left\{ E\left[P^{m_n}(X_j)' \tilde{\boldsymbol{\beta}}_j - f_j^*(X_j) \right]^2 \right\}^{1/2} = O\left(m_n^{-d} \right). \tag{A.4}$$

We rewrite Y as

$$\begin{aligned} Y &= \sum_{j=1}^{p^*} P^{m_n}(X_j)' \tilde{\boldsymbol{\beta}}_j + \sum_{j=1}^{p^*} \left[f_j^*(X_j) - P^{m_n}(X_j)' \tilde{\boldsymbol{\beta}}_j \right] + \varepsilon \\ &\equiv \sum_{j=1}^{p^*} P^{m_n}(X_j)' \tilde{\boldsymbol{\beta}}_j + R_n + \varepsilon. \end{aligned} \tag{A.5}$$

Equation (A.4) implies $R_n = O_P\left(m_n^{-d}\right)$ for a fixed p^*. Then the model looks like a linear model with an approximation bias term R_n and an error term ε.

Now we have a closer look at $\theta_{l,\boldsymbol{Z}}$, which measures the impact of X_l on Y after controlling $\boldsymbol{Z}$. If we can ignore the approximation bias, a finite sample approximation to the term inside $\theta_{l,\boldsymbol{Z}}$, namely,

$$E\left\{ Y - P^{m_n}(\boldsymbol{Z})' \left\{ E\left[P^{m_n}(\boldsymbol{Z}) P^{m_n}(\boldsymbol{Z})' \right] \right\}^{-1} E\left[P^{m_n}(\boldsymbol{Z}) Y \right] \middle| X_l \right\},$$

is

$$\begin{aligned} & P^{m_n}(X_l)' \Phi_{X_l}^{-1} \cdot E\left\{ P^{m_n}(X_l) \left\{ Y - P^{m_n}(\boldsymbol{Z})' \Phi_{\boldsymbol{Z}}^{-1} E\left[P^{m_n}(\boldsymbol{Z}) Y \right] \right\} \right\} \\ & \propto m_n \cdot P^{m_n}(X_l)' E\left\{ P^{m_n}(X_l) \left\{ Y - P^{m_n}(\boldsymbol{Z})' \Phi_{\boldsymbol{Z}}^{-1} E\left[P^{m_n}(\boldsymbol{Z}) Y \right] \right\} \right\}, \end{aligned} \tag{A.6}$$

where $\Phi_{X_l} \equiv E[P^{m_n}(X_l) P^{m_n}(X_l)']$, $\Phi_{\boldsymbol{Z}} \equiv E[P^{m_n}(\boldsymbol{Z}) P^{m_n}(\boldsymbol{Z})']$, and the last line holds by Lemma A.3. We let

$$\boldsymbol{\eta}_{l,\boldsymbol{Z}} \equiv E\left\{ P^{m_n}(X_l) \left\{ Y - P^{m_n}(\boldsymbol{Z})' \Phi_{\boldsymbol{Z}}^{-1} E\left[P^{m_n}(\boldsymbol{Z}) Y \right] \right\} \right\},$$

which appears as the numerator in the population coefficient of the regression of $Y - P^{m_n}(\boldsymbol{Z})' \Phi_{\boldsymbol{Z}}^{-1} \cdot E\left[P^{m_n}(\boldsymbol{Z}) Y \right]$ on $P^{m_n}(X_l)$. Then the last term in equation (A.6) becomes $m_n P^{m_n}(X_l)' \boldsymbol{\eta}_{l,\boldsymbol{Z}}$. As a result,

$$\theta_{l,\boldsymbol{Z}} \propto \left\{ E\left[\left(m_n P^{m_n}(X_l)' \boldsymbol{\eta}_{l,\boldsymbol{Z}} \right)^2 \right] \right\}^{1/2} = \left\{ m_n^2 \boldsymbol{\eta}_{l,\boldsymbol{Z}}' \Phi_{X_l} \boldsymbol{\eta}_{l,\boldsymbol{Z}} \right\}^{1/2} \propto m_n^{1/2} \left\| \boldsymbol{\eta}_{l,\boldsymbol{Z}} \right\| \tag{A.7}$$

by Lemma A.3. Note $\theta_{l,\boldsymbol{Z}}$ measures the impact of X_l on Y after controlling $\boldsymbol{Z}$. $m_n^{1/2} \left\| \boldsymbol{\eta}_{l,\boldsymbol{Z}} \right\|$ can be equivalently used to measure this strength.

To facilitate the analysis of $\boldsymbol{\eta}_{l,\boldsymbol{Z}}$, let $\Phi_{X_l \boldsymbol{Z}} \equiv E[P^{m_n}(X_l) P^{m_n}(\boldsymbol{Z})']$. By the orthogonality between the error term and regressor in the least squares projection, we have

$$\begin{aligned} \boldsymbol{\eta}_{l,\boldsymbol{Z}} &= E\left\{ \left[P^{m_n}(X_l) - \Phi_{X_l \boldsymbol{Z}} \Phi_{\boldsymbol{Z}}^{-1} P^{m_n}(\boldsymbol{Z}) \right] \left\{ Y - P^{m_n}(\boldsymbol{Z})' \Phi_{\boldsymbol{Z}}^{-1} E\left[P^{m_n}(\boldsymbol{Z}) Y \right] \right\} \right\} \\ &= E\left\{ \left[P^{m_n}(X_l) - \boldsymbol{\gamma}_{X_l,\boldsymbol{Z}}' P^{m_n}(\boldsymbol{Z}) \right] \cdot \left[Y - \boldsymbol{\gamma}_{Y,\boldsymbol{Z}}' P^{m_n}(\boldsymbol{Z}) \right] \right\} \\ &= E\left(\boldsymbol{U}_{X_l,\boldsymbol{Z}} \cdot U_{Y,\boldsymbol{Z}} \right), \end{aligned} \tag{A.8}$$

where $\boldsymbol{\gamma}_{X_l,\boldsymbol{Z}} \equiv \Phi_{\boldsymbol{Z}}^{-1} \Phi_{X_l \boldsymbol{Z}}'$, $\boldsymbol{\gamma}_{Y,\boldsymbol{Z}} \equiv \Phi_{\boldsymbol{Z}}^{-1} E\left[P^{m_n}(\boldsymbol{Z}) Y \right]$, and $\boldsymbol{U}_{X_l,\boldsymbol{Z}}$ and $U_{Y,\boldsymbol{Z}}$ denote the respective projection errors:

$$\boldsymbol{U}_{X_l,\boldsymbol{Z}} \equiv P^{m_n}(X_l) - \boldsymbol{\gamma}_{X_l,\boldsymbol{Z}}' P^{m_n}(\boldsymbol{Z}), \text{ and } U_{Y,\boldsymbol{Z}} \equiv Y - \boldsymbol{\gamma}_{Y,\boldsymbol{Z}}' P^{m_n}(\boldsymbol{Z}). \tag{A.9}$$

Note that $\boldsymbol{U}_{X_l,\boldsymbol{Z}}$ is $m_n \times 1$ and $U_{Y,\boldsymbol{Z}}$ is a scalar. Thus, $\boldsymbol{\eta}_{l,\boldsymbol{Z}}$ measures the covariance between $P^{m_n}(X_l)$ and Y after controlling $P^{m_n}(\boldsymbol{Z})$.

At the sample level, we use the following notations. Let $\boldsymbol{z}_i$ denote the i-th observation for $\boldsymbol{Z}$. We let $\boldsymbol{u}_{X_l,\boldsymbol{Z}} \equiv (\boldsymbol{u}_{X_l,\boldsymbol{Z},1}, \ldots, \boldsymbol{u}_{X_l,\boldsymbol{Z},n})'$ and $\boldsymbol{u}_{Y,\boldsymbol{Z}} \equiv (u_{Y,\boldsymbol{Z},1}, \ldots, u_{Y,\boldsymbol{Z},n})'$, where $\boldsymbol{u}_{X_l,\boldsymbol{Z},i} = P^{m_n}(x_{li}) - \boldsymbol{\gamma}'_{X_l,\boldsymbol{Z}} P^{m_n}(\boldsymbol{z}_i)$ and $u_{Y,\boldsymbol{Z},i} = y_i - \boldsymbol{\gamma}'_{Y,\boldsymbol{Z}} P^{m_n}(\boldsymbol{z}_i)$. The connection between $\hat{\mathcal{X}}_{l,\boldsymbol{Z}}$ and $\boldsymbol{\eta}_{l,\boldsymbol{Z}}$ can be seen from the following. Noting that

$$\boldsymbol{y} = \mathbb{Z}\boldsymbol{\gamma}_{Y,\boldsymbol{Z}} + \boldsymbol{u}_{Y,\boldsymbol{Z}} \text{ and } \mathbb{X}_l = \mathbb{Z}\boldsymbol{\gamma}_{X_l,\boldsymbol{Z}} + \boldsymbol{u}_{X_l,\boldsymbol{Z}},$$

we have

$$\begin{aligned}
\hat{\mathcal{X}}_{l,\boldsymbol{Z}} &= \left(\boldsymbol{u}'_{Y,\boldsymbol{Z}} M_{\mathbb{Z}} \boldsymbol{u}_{X_l,\boldsymbol{Z}}\right) \left(\hat{\sigma}^2_{l,\boldsymbol{Z}} \mathbb{X}'_l M_{\mathbb{Z}} \mathbb{X}_l\right)^{-1} \left(\boldsymbol{u}'_{X_l,\boldsymbol{Z}} M_{\mathbb{Z}} \boldsymbol{u}_{Y,\boldsymbol{Z}}\right) \\
&= \left[\left(\boldsymbol{u}'_{X_l,\boldsymbol{Z}} \boldsymbol{u}_{Y,\boldsymbol{Z}} - n\boldsymbol{\eta}_{l,\boldsymbol{Z}} - \boldsymbol{u}'_{X_l,\boldsymbol{Z}} Q_{\mathbb{Z}} \boldsymbol{u}_{Y,\boldsymbol{Z}}\right) + n\boldsymbol{\eta}_{l,\boldsymbol{Z}}\right]' \left(\hat{\sigma}^2_{l,\boldsymbol{Z}} \mathbb{X}'_l M_{\mathbb{Z}} \mathbb{X}_l\right)^{-1} \\
&\cdot \left[\left(\boldsymbol{u}'_{X_l,\boldsymbol{Z}} \boldsymbol{u}_{Y,\boldsymbol{Z}} - n\boldsymbol{\eta}_{l,\boldsymbol{Z}} - \boldsymbol{u}'_{X_l,\boldsymbol{Z}} Q_{\mathbb{Z}} \boldsymbol{u}_{Y,\boldsymbol{Z}}\right) + n\boldsymbol{\eta}_{l,\boldsymbol{Z}}\right] \\
&= n^2 \boldsymbol{\eta}'_{l,\boldsymbol{Z}} \left(\hat{\sigma}^2_{l,\boldsymbol{Z}} \mathbb{X}'_l M_{\mathbb{Z}} \mathbb{X}_l\right)^{-1} \boldsymbol{\eta}_{l,\boldsymbol{Z}} + 2n\boldsymbol{\eta}'_{l,\boldsymbol{Z}} \left(\hat{\sigma}^2_{l,\boldsymbol{Z}} \mathbb{X}'_l M_{\mathbb{Z}} \mathbb{X}_l\right)^{-1} \tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}} + \tilde{\boldsymbol{u}}'_{l,\boldsymbol{Z}} \left(\hat{\sigma}^2_{l,\boldsymbol{Z}} \mathbb{X}'_l M_{\mathbb{Z}} \mathbb{X}_l\right)^{-1} \tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}},
\end{aligned} \tag{A.10}$$

where $Q_{\mathbb{Z}} = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'$ and

$$\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}} \equiv \boldsymbol{u}'_{X_l,\boldsymbol{Z}} \boldsymbol{u}_{Y,\boldsymbol{Z}} - n\boldsymbol{\eta}_{l,\boldsymbol{Z}} - \boldsymbol{u}'_{X_l,\boldsymbol{Z}} Q_{\mathbb{Z}} \boldsymbol{u}_{Y,\boldsymbol{Z}}. \tag{A.11}$$

behaves like the counterpart of $\mathbb{X}'_l \boldsymbol{u}_l$ in equation (A.3). To see the last point, first note that

$$E\left[\boldsymbol{u}'_{X_l,\boldsymbol{Z}} \boldsymbol{u}_{Y,\boldsymbol{Z}} - n\boldsymbol{\eta}_{l,\boldsymbol{Z}}\right] = 0.$$

Next, it is easy to show that $\boldsymbol{u}'_{X_l,\boldsymbol{Z}} Q_{\mathbb{Z}} \boldsymbol{u}_{Y,\boldsymbol{Z}} = n^{-1/2} \boldsymbol{u}'_{X_l,\boldsymbol{Z}} \mathbb{Z} \left(n^{-1}\mathbb{Z}'\mathbb{Z}\right)^{-1} n^{-1/2} \mathbb{Z}' \boldsymbol{u}_{Y,\boldsymbol{Z}}$ is of smaller order than $\boldsymbol{u}'_{X_l,\boldsymbol{Z}} \boldsymbol{u}_{Y,\boldsymbol{Z}} - n\boldsymbol{\eta}_{l,\boldsymbol{Z}}$ because $E\left[P^{m_n}(\boldsymbol{Z})\, \boldsymbol{U}'_{X_l,\boldsymbol{Z}}\right] = \boldsymbol{0}$ and $E\left[P^{m_n}(\boldsymbol{Z})\, U_{Y,\boldsymbol{Z}}\right] = \boldsymbol{0}$ by the definitions in equation (A.9). The first term in (A.11) is the dominant term of $\hat{\mathcal{X}}_{l,\boldsymbol{Z}}$ if $\boldsymbol{\eta}_{l,\boldsymbol{Z}}$ is large enough.

The analysis of $\hat{\mathcal{X}}_l(\boldsymbol{Z})$ is based on the decomposition in equation (A.10). The basic idea is similar to the one for analyzing $\hat{\mathcal{X}}_l$ in equation (A.3): we show $\hat{\sigma}^2_{l,\boldsymbol{Z}}$, $n^{-1}\mathbb{X}'_l M_{\mathbb{Z}} \mathbb{X}_l$, and $n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}$ are bounded by some value or rate with very high probability, and as a result, $\mathbf{1}\left(\hat{\mathcal{X}}_l(\boldsymbol{Z}) > \varsigma_n\right)$ is determined by the strength of $\theta_{l,\boldsymbol{Z}}$ (or equivalently, $m_n^{1/2}\left\|\boldsymbol{\eta}_{l,\boldsymbol{Z}}\right\|$) with very high probability.

A.4 Technical Lemmas

We frequently use the inequalities presented in Remark 1, Lemma A.1, and Lemma A.2.

Remark 1 (Some inequalities) We may apply the following inequalities in the proofs directly without referring them back to here. First

$$\Pr(X_1 + X_2 \geq C) \leq \Pr(X_1 \geq \pi C) + \Pr(X_2 \geq (1-\pi)C)$$

for any constant π. That is due to the fact that $\{X_1 + X_2 \geq C\} \subseteq \{X_1 \geq \pi C\} \cup \{X_1 \geq (1-\pi) C\}$. Similarly, we have

$$\Pr\left(\sum_{i=1}^{n} X_i \geq C\right) \leq \sum_{i=1}^{n} \Pr\left(X_i \geq Cn^{-1}\right) \text{ and } \Pr\left(\max_{1\leq i\leq n} X_i \geq C\right) \leq \sum_{i=1}^{n} \Pr\left(X_i \geq C\right).$$

For any positive random variables X_1 and X_2 and positive constants C_1 and C_2,

$$\Pr\left(X_1 \cdot X_2 \geq C_1\right) \leq \Pr\left(X_1 \geq C_1/C_2\right) + \Pr\left(X_2 \geq C_2\right),$$

due to the fact that $\{X_1 \cdot X_2 \geq C_1\} \subseteq \{X_1 \geq C_1/C_2\} \cup \{X_2 \geq C_2\}$.

Lemma A.1 (Bernstein's Inequality) *For an i.i.d. series* $\{z_i\}_{i=1}^{\infty}$ *with zero mean and bounded support* $[-M, M]$, *we have*

$$\Pr\left(\left|\sum_{i=1}^{n} z_i\right| > x\right) \leq 2\exp\left\{-x^2/\left[2\left(v + Mx/3\right)\right]\right\},$$

for any $v \geq \mathrm{Var}(\sum_{i=1}^{n} z_i)$.

The above lemma is Lemma 2.2.9 in van der Vaart and Wellner (1996).

Lemma A.2 *Suppose that for an i.i.d. series* $\{z_i\}_{i=1}^{\infty}$, $\Pr\left(|z_i| > x\right) \leq C_1 \exp\left(-C_2 x^{\alpha}\right)$ *holds for all* $x > 0$ *for some* $C_1, C_2, C_3 > 0$. *Further* $E\left(z_i\right) = 0$ *and* $E\left(z_i^2\right) = \sigma_z^2$. *Let* $\{v_n\}_{n=1}^{\infty}$ *be a deterministic series such that* $v_n \propto n^{\lambda}$. *Then (i) if* $0 < \lambda \leq (\alpha+1)/(\alpha+2)$, *we have*

$$\Pr\left(\left|\sum_{i=1}^{n} z_i\right| > v_n\right) \leq \exp\left[-\left(1-\pi\right)^2 v_n^2 \Big/ \left(2n\sigma_z^2\right)\right] \text{ for any } \pi \in (0,1);$$

(ii) if $\lambda \geq (\alpha+1)/(\alpha+2)$, *we have*

$$\Pr\left(\left|\sum_{i=1}^{n} z_i\right| > v_n\right) \leq \exp\left(-C_4 v_n^{\alpha/(\alpha+1)}\right) \text{ for some } C_4 > 0.$$

This lemma is Lemma A3 in CKP.

Lemma A.3 *Recall that* $\Phi_{X_l} = E[P^{m_n}(X_l) P^{m_n}(X_l)']$. *Suppose that Assumption 3 holds. Then there exist some positive constants* B_1, B_2, B_3, *and* B_4 *such that*

$$\begin{aligned} B_1 m_n^{-1} &\leq \lambda_{\min}\left(\Phi_{X_l}\right) \leq \lambda_{\max}\left(\Phi_{X_l}\right) \leq B_2 m_n^{-1}, \quad \text{(A.12)} \\ B_3 m_n^{-1} &\leq E\left[\phi_k\left(X_l\right)^2\right] \leq B_4 m_n^{-1}, \text{ and} \\ E\left[\left|\phi_k\left(X_l\right)\right|\right] &\propto m_n^{-1}, \end{aligned}$$

for all k *and* l. *For any* $f_{nl}\left(X_l\right) = P^{m_n}\left(X_l\right)' \boldsymbol{\beta}$ *where* X_l *satisfies Assumption 3, we have*

$$\left\{E\left[f_{nl}\left(X_l\right)^2\right]\right\}^{1/2} \propto m_n^{-1/2} \left\|\boldsymbol{\beta}\right\|.$$

For the first and second equations in the above lemma, see Lemma 6.2 in Zhou, Shen and Wolfe (1998). The third equation is a direct result of those on pp. 91 and 133 in de Boor (2001). The last equation follows from (A.12) and the fact

$$\left\{E\left[f_{nl}\left(X_l\right)^2\right]\right\}^{1/2} = \left\{\boldsymbol{\beta}' E\left[P^{m_n}\left(X_l\right) P^{m_n}\left(X_l\right)'\right] \boldsymbol{\beta}\right\}^{1/2}, \text{ and}$$
$$\left[\lambda_{\min}\left(\Phi_{X_l}\right)\right]^{1/2} \|\boldsymbol{\beta}\| \leq \left\{\boldsymbol{\beta}' E\left[P^{m_n}\left(X_l\right) P^{m_n}\left(X_l\right)'\right] \boldsymbol{\beta}\right\}^{1/2} \leq \left[\lambda_{\max}\left(\Phi_{X_l}\right)\right]^{1/2} \|\boldsymbol{\beta}\|.$$

Lemma A.4 *Suppose that Assumptions 1 and 3 hold. Then*

$$\Pr\left(\left|\left\|\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\right\| - \left\|\Phi_{X_l}^{-1}\right\|\right| > \left\|\Phi_{X_l}^{-1}\right\| \Big/ 2\right) \leq 2m_n^2 \exp\left\{-C_1 n m_n^{-3}\right\} \tag{A.13}$$

for some positive constant C_1. If Assumption 7 also holds, then

$$\Pr\left(\left|\left\|\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\right\| - \left\|\Phi_{X_l}^{-1}\right\|\right| > \left\|\Phi_{X_l}^{-1}\right\| \Big/ 2\right) \leq C_2 \exp\left\{-C_3 n^{C_4}\right\}$$

for some positive C_2, C_3, and C_4.

Lemma A.5 *Suppose that Assumptions 1, 3, 4, 6, and 7 hold. Let $\sigma_l^2 = E\left(U_l^2\right)$, $\omega_l^4 = E\left(U_l^4\right) - \left[E\left(U_l^2\right)\right]^2$, $u_{li} = y_i - f_{nl}\left(x_{li}\right)$, $\hat{u}_{li} = y_i - P^{m_n}\left(x_{li}\right)' \hat{\boldsymbol{\beta}}_l$, and $\hat{\sigma}_l^2 = n^{-1}\sum_{i=1}^n \hat{u}_{li}^2$. Then, for $v_n \propto n^\lambda$ with $\lambda > 1/2$,*

$$\Pr\left(\left|\hat{\sigma}_l^2 - \sigma_l^2\right| \geq v_n/n\right) \leq C_1 \exp\left(-C_2 n^{C_3}\right)$$

holds for some C_1, C_2, and $C_3 > 0$.

Lemma A.6 *Suppose Assumptions 1, 3, 4, 6, and 7 hold.*

(i) If $v_n \lesssim n^{s/(s+2)} m_n^2$, then

$$\Pr\left(\left|\boldsymbol{u}_l'\mathbb{X}_l\left(\hat{\sigma}_l^2\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right| \geq \frac{4}{3}\sigma_l^2 v_n\right) \leq \exp\left(-C_1 m_n^{-1} v_n + \log m_n\right) + C_2 \exp\left(-C_3 n^{C_4}\right)$$

for some positive constants C_1, C_2, C_3, and C_4.

(ii) If $v_n \gtrsim n^{s/(s+2)} m_n^2$, then

$$\Pr\left(\left|\boldsymbol{u}_l'\mathbb{X}_l\left(\hat{\sigma}_l^2\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right| \geq \frac{4}{3}\sigma_l^2 v_n\right) \leq C_5 \exp\left(-C_6 n^{C_7}\right)$$

for some positive constants C_5, C_6, and C_7.

Lemma A.7 *Suppose that Assumptions 3, 5', and 7 hold and that we pre-select $\mathbf{Z} = \left(Z_1, \ldots, Z_{\iota_n}\right)'$ in previous stages and all Zs are either signals or pseudo-signals. Assume that b signals, without loss of generality say $\mathbf{X}_1^b \equiv \left(X_1, \ldots, X_b\right)'$, are not in the list of the pre-selected variables. Let*

$$\boldsymbol{\eta}_1^b = E\left\{P^{m_n}\left(\mathbf{X}_1^b\right)\left\{Y - P^{m_n}\left(\mathbf{Z}\right)'\left\{E\left[P^{m_n}\left(\mathbf{Z}\right)P^{m_n}\left(\mathbf{Z}\right)'\right]\right\}^{-1} E\left[P^{m_n}\left(\mathbf{Z}\right)Y\right]\right\}\right\}$$

If there exist a j, $1 \leq j \leq b$, *such that* $\left\{E\left[f_{nj}^{*}\left(X_{j}\right)^{2}\right]\right\}^{1/2} \gtrsim \kappa_{n} \log\left(m_{n}\right)^{1/2}\left(m_{n}/n\right)^{1/2}$ *for some* $\kappa_{n} \to \infty$, *we have*

$$\left\|\boldsymbol{\eta}_{1}^{b}\right\| \gtrsim \kappa_{n} \log\left(m_{n}\right)^{1/2} n^{-1/2}.$$

Lemma A.8 *Suppose Assumptions 1, 3, and 10 hold. Further suppose all covariates in* $\mathbf{Z}$ $(\iota_n \times 1)$ *are either signals or pseudo signals. For a random variable* X_l *that is not included in* $\mathbf{Z}$, *it holds that*

$$\Pr\left(\left|\left\|\left(n^{-1}\begin{pmatrix}\mathbb{Z}'\\ \mathbb{X}_l'\end{pmatrix}\begin{pmatrix}\mathbb{Z} & \mathbb{X}_l\end{pmatrix}\right)^{-1}\right\| - \left\|\Phi^{-1}\right\|\right| > \left\|\Phi^{-1}\right\|/2\right) \leq 2m_n^2\iota_n^2\exp\left\{-C_1 n m_n^{-3}\iota_n^{-1}\right\}$$

for some positive constant C_1, *where* $\Phi = \begin{bmatrix}\Phi_{\boldsymbol{Z}} & \Phi_{\boldsymbol{Z}X_l}\\ \Phi_{\boldsymbol{Z}X_l}' & \Phi_{X_l}\end{bmatrix}$, $\Phi_{\boldsymbol{Z}X_l} = E\left[P^{m_n}(\mathbf{Z})'P^{m_n}(X_l)\right]$, *and recall that* $\Phi_{\boldsymbol{Z}} = E[P^{m_n}(\mathbf{Z})\,P^{m_n}(\mathbf{Z})']$ *and* $\Phi_{X_l} = E\left[P^{m_n}(X_l)\,P^{m_n}(X_l)'\right]$. *If in addition Assumptions 2' and 7 hold, the right-hand side of the inequality can be replaced by* $C_2\exp\left(-C_3 n^{C_4}\right)$.

Lemma A.9 *Suppose Assumptions 1, 2', 3, 4, 6, 7, and 10 hold. Further suppose all covariates in* $\mathbf{Z}$ $(\iota_n \times 1)$ *are either signals or pseudo signals. For a random variable* X_l *that is not included in* $\mathbf{Z}$, *then*

$$\Pr\left(\lambda_{\max}\left\{\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2 n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\} \geq 2\sigma^{-2}B_{X1}^{-1}m_n\right) \leq C_1\exp(-C_2 n^{C_3}) \text{ and}$$
$$\Pr\left(\lambda_{\min}\left\{\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2 n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\} \leq \frac{1}{4}\sigma^{-2}B_{X2}^{-1}m_n\right) \leq C_4\exp(-C_5 n^{C_6}),$$

for some positive $C_1, \ldots, C_6$, *where* σ^2 *is defined in Lemma A.11.*

Lemma A.10 *Suppose Assumptions 1, 2', 3, 4, 5', 6, 7, and 10 hold. Further suppose all covariates in* $\mathbf{Z}$ $(\iota_n \times 1)$ *are either signals or pseudo signals. For a random variable* X_l *that is not included in* $\mathbf{Z}$,*then*

$$\Pr\left(\left\|n^{-1/2}\left(\boldsymbol{u}_{X_l,\boldsymbol{Z}}'\boldsymbol{u}_{Y,\boldsymbol{Z}} - \boldsymbol{\eta}_{l,\boldsymbol{Z}}\right)\right\|^2 \geq C_1 m_n^{-1}v_n\right)$$
$$\leq \begin{cases} 2m_n\exp\left\{-n m_n^{-2}v_n/[C_2(n m_n^{-1} + \iota_n n^{1/2}v_n^{1/2})]\right\} + m_n\exp\left(-C_3 m_n^{-1}v_n\right) & \text{if } v_n \lesssim n^{s/(s+2)}m_n^2 \\ 2m_n\exp\left\{-n m_n^{-2}v_n/[C_2(n m_n^{-1} + \iota_n n^{1/2}v_n^{1/2})]\right\} + C_4\exp\left(-C_5 n^{C_6}\right) & \text{if } v_n \gtrsim n^{s/(s+2)}m_n^2 \end{cases},$$

and

$$\Pr\left(\left\|n^{-1/2}\boldsymbol{u}_{X_l,\boldsymbol{Z}}'\mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}\right\|^2 \geq C_1 m_n^{-1}v_n\right)$$
$$\leq \begin{cases} C_{13}\exp\left(-C_{14}n^{C_{15}}\right) + m_n^2\iota_n\exp\left(-C_{16}n^{1/2}m_n^{-1/6}\iota_n^{-1}\left(m_n^{-1}v_n\right)^{1/2}\right) & \text{if } v_n \lesssim n^{(s-2)/[4(s+2)]}m_n^{5/6}\iota_n^2 \\ C_{17}\exp\left(-C_{18}n^{C_{19}}\right) & \text{if } v_n \gtrsim n^{(s-2)/[4(s+2)]}m_n^{5/6}\iota_n^2 \end{cases},$$

for some positive $C_2, C_3, C_4, C_5,$ *and* C_6*. If we set* $v_n = \varsigma_n$*, then*

$$\Pr\left(\left\|n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right\|^2 \geq C_1 m_n^{-1}\varsigma_n\right) \leq n^{-M} + C_{20}\exp\left(-C_{21}n^{C_{22}}\right)$$

for any large positive constant M *and some positive constants* $C_{20}, C_{21},$ *and* C_{22}*.*

Lemma A.11 *Suppose Assumptions 1, 2', 3, 4, 5', 7, and 10 hold. Suppose all covariates in* $\boldsymbol{Z}$ *(*$\iota_n \times 1$*) are either signals or pseudo signals and* X_l *is a variable that is not included in* $\boldsymbol{Z}$*. Define*

$$U = Y - \begin{pmatrix} P^{m_n}(\boldsymbol{Z}) \\ P^{m_n}(X_l) \end{pmatrix}' \Phi^{-1} E\left[\begin{pmatrix} P^{m_n}(\boldsymbol{Z}) \\ P^{m_n}(X_l) \end{pmatrix} Y\right],$$

Let $\sigma^2 = E\left(U^2\right), \omega^4 = E\left(U^4\right) - \left[E\left(U^2\right)\right]^2$*. Then, for* $v_n \propto n^\lambda$ *with* $\lambda > 1/2$*, we have*

$$\Pr\left(\left|\hat{\sigma}^2_{l,\boldsymbol{Z}} - \sigma^2\right| \geq v_n/n\right) \leq C_1\exp\left(-C_2 n^{C_3}\right) \text{ for some } C_1, C_2, C_3 > 0.$$

Lemma A.12 *Suppose Assumptions 1, 2', 3, 4, 6, 7, and 10 hold. Further suppose all covariates in* $\boldsymbol{Z}$ *(*$\iota_n \times 1$*) are either signals or pseudo signals. For a random variable* X_l *that is not included in* $\boldsymbol{Z}$*, we have*

$$\Pr\left(\tilde{\boldsymbol{u}}'_{l,\boldsymbol{Z}}\left(\hat{\sigma}^2_{l,\boldsymbol{Z}}\mathbb{X}'_l M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}} \geq \varsigma_n\right) \leq \exp\left(-C_1 m_n^{-1}\varsigma_n + \log m_n\right) + C_2\exp(-C_3 n^{C_4}).$$

Lemma A.13 *Suppose Assumptions 1, 2', 3, 4, 5', 6, 7, 8, and 10 hold. Then*

$$\Pr\left(\mathcal{D}_k\right) \geq 1 - kn^{-M_1} - kC_4\exp\left(-C_5 n^{C_6}\right)$$

for any finite fixed positive integer k*, any fixed large positive constant* M_1*, and some positive constants* $C_4, C_5,$ *and* C_6*. When* k *is fixed or divergent to infinity at a rate no faster than* n^a *for some* $a > 0$*, we can write*

$$\Pr\left(\mathcal{D}_k\right) \geq 1 - n^{-M_2} - C_7\exp\left(-C_8 n^{C_9}\right) \text{ for any } k \leq n^a,$$

for some large positive constant M_2 *and some positive constants* $C_7, C_8,$ *and* C_9*.*

B Proofs of the Main Results

Proof of Proposition 2.1. To see the intuition, we begin with the case where $\theta_l = 0$. When $\theta_l = 0$, $\|\boldsymbol{\beta}_l\| = 0$ and $\hat{\mathcal{X}}_l = \boldsymbol{u}'_l\mathbb{X}_l\left(\hat{\sigma}^2_l\mathbb{X}'_l\mathbb{X}_l\right)^{-1}\mathbb{X}'_l\boldsymbol{u}_l$. Then,

$$\begin{aligned}\Pr\left(\hat{\mathcal{X}}_l \geq \varsigma_n\right) &\leq \Pr\left(\boldsymbol{u}'_l\mathbb{X}_l\left(\hat{\sigma}^2_l\mathbb{X}'_l\mathbb{X}_l\right)^{-1}\mathbb{X}'_l\boldsymbol{u}_l \geq \varsigma_n\right) \\ &= \Pr\left(\boldsymbol{u}'_l\mathbb{X}_l\left(\hat{\sigma}^2_l\mathbb{X}'_l\mathbb{X}_l\right)^{-1}\mathbb{X}'_l\boldsymbol{u}_l \geq \frac{4}{3}\sigma^2_l\left(\frac{3}{4}\sigma^{-2}_l\varsigma_n\right)\right)\end{aligned}$$

$$\leq \exp\left(-C_1 m_n^{-1}\varsigma_n + \log m_n\right) + C_2 \exp\left(-C_3 n^{C_4}\right)$$

for some positive constants C_1, C_2, and C_3, where the last line holds by applying the first part of Lemma A.6 with $v_n = \frac{3}{4}\sigma_l^{-2}\varsigma_n$. Since $\varsigma_n \propto \kappa_n \log(m_n) m_n$, $C_1 m_n^{-1}\varsigma_n - \log m_n \gg M \log m_n$ for any fixed positive M. This shows the first part of the proposition when $\theta_l = 0$.

Note by Lemma A.3

$$\left\{E\left[f_{nl}(X_l)^2\right]\right\}^{1/2} \propto m_n^{-1/2} \|\boldsymbol{\beta}_l\|.$$

If $\theta_l \lesssim \log(m_n)^{1/2}(m_n/n)^{1/2}$, by the triangle inequality, equation (A.2) and Assumption 7,

$$\begin{aligned}\left\{E\left[f_{nl}(X_l)^2\right]\right\}^{1/2} &\leq \left\{E\left[f_l(X_l)^2\right]\right\}^{1/2} + \left\{E\left[|f_{nl}(X_l) - f_l(X_l)|^2\right]\right\}^{1/2} \\ &\lesssim \log(m_n)^{1/2}(m_n/n)^{1/2} + C_1 m_n^{-d} \lesssim \log(m_n)^{1/2}(m_n/n)^{1/2},\end{aligned}$$

because the approximation bias is asymptotically negligible under Assumption 7. Therefore, $\|\boldsymbol{\beta}_l\| \lesssim \log(m_n)^{1/2} m_n n^{-1/2}$. Then,

$$\begin{aligned}&\Pr\left(\hat{\mathcal{X}}_l \geq \varsigma_n\right) \\ &= \Pr\left(\left(n^{1/2}\boldsymbol{\beta}_l + n^{1/2}\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right)'\left(\hat{\sigma}_l^{-2}n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)\left(n^{1/2}\boldsymbol{\beta}_l + n^{1/2}\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right) \geq \varsigma_n\right) \\ &\leq \Pr\left(\left\|n^{1/2}\boldsymbol{\beta}_l + n^{1/2}\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right\|^2 \hat{\sigma}_l^{-2}\lambda_{\max}\left\{n^{-1}\mathbb{X}_l'\mathbb{X}_l\right\} \geq \varsigma_n\right) \\ &\leq \Pr\left(\left\|n^{1/2}\boldsymbol{\beta}_l + n^{1/2}\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right\|^2 \geq \frac{1}{3}\sigma_l^2 B_2^{-1} m_n\varsigma_n\right) + \Pr\left(\hat{\sigma}_l^{-2} \geq 2\sigma_l^{-2}\right) \\ &+ \Pr\left(\lambda_{\max}\left\{\left(\mathbb{X}_l'\mathbb{X}_l\right)\right\} \geq \frac{3}{2}B_2 m_n^{-1}\right) \\ &\equiv A_1 + A_2 + A_3.\end{aligned}$$

By Lemma A.5 and Lemma A.4, $A_l = \frac{1}{4}C_2\exp\left(-C_3 n^{C_4}\right)$ for some constants C_2, C_3 and C_4 and $l = 2, 3$. For A_1, we have with $C \equiv \frac{1}{3}\sigma_l^2 B_2^{-1}$,

$$\begin{aligned}A_1 &\leq \Pr\left(\left\|n^{1/2}\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right\| + \left\|n^{1/2}\boldsymbol{\beta}_l\right\| \geq C m_n^{1/2}\varsigma_n^{1/2}\right) \\ &= \Pr\left(\left\|\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1} n^{-1/2}\mathbb{X}_l'\boldsymbol{u}_l\right\| \geq \frac{1}{2}C m_n^{1/2}\varsigma_n^{1/2} + \frac{1}{2}C m_n^{1/2}\varsigma_n^{1/2} - \left\|n^{1/2}\boldsymbol{\beta}_l\right\|\right) \\ &\leq \Pr\left(\left\|\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1} n^{-1/2}\mathbb{X}_l'\boldsymbol{u}_l\right\| \geq \frac{1}{2}C m_n^{1/2}\varsigma_n^{1/2}\right) \\ &\leq \Pr\left(\left\|n^{-1/2}\mathbb{X}_l'\boldsymbol{u}_l\right\| \geq \frac{1}{2}C m_n^{-1/2}\varsigma_n^{1/2}\right) + \Pr\left(\left\|\lambda_{\max}\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\right\| \geq \frac{3}{2}B_1^{-1} m_n\right) \\ &\leq n^{-M} + \frac{1}{2}C_2\exp\left(-C_3 n^{C_4}\right)\end{aligned}$$

for any arbitrarily large constant M and some constants C_2, C_3 and C_4, where for the fourth line holds by the fact that $m_n^{1/2}\varsigma_n^{1/2} = \kappa_n^{1/2}\log(m_n)^{1/2} m_n \gg \left\|n^{1/2}\boldsymbol{\beta}_l\right\|$, and the last line holds by equation (S1.14), the fact that $C_1 m_n^{-1}\varsigma_n - \log m_n \gg M\log m_n$, and Lemma A.4. In sum, we have $\Pr\left(\hat{\mathcal{X}}_l \geq \varsigma_n\right) \leq n^{-M} + C_2\exp\left(-C_3 n^{C_4}\right)$.

We turn to the second part of the proposition. If $\theta_l \gtrsim \kappa_n \log(m_n)^{1/2}(m_n/n)^{1/2}$, by equation (A.2) and Assumption 7,

$$\left\{E\left[f_{nl}(X_l)^2\right]\right\}^{1/2} \gtrsim \kappa_n \log(m_n)^{1/2}(m_n/n)^{1/2},$$

where we use the fact that the approximation bias is asymptotically negligible under Assumption 7. Therefore, $\|\boldsymbol{\beta}_l\| \gtrsim \kappa_n \log(m_n)^{1/2} m_n n^{-1/2}$.

We bound $\Pr\left(\hat{\mathcal{X}}_l < \varsigma_n\right)$ as follows:

$$\begin{aligned}
&\Pr\left(\hat{\mathcal{X}}_l < \varsigma_n\right)\\
&= \Pr\left(\left(n^{1/2}\boldsymbol{\beta}_l + n^{1/2}\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right)'\left(\hat{\sigma}_l^{-2}n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)\left(n^{1/2}\boldsymbol{\beta}_l + n^{1/2}\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right) < \varsigma_n\right)\\
&\leq \Pr\left(\left\|n^{1/2}\boldsymbol{\beta}_l + n^{1/2}\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right\|^2 \lambda_{\min}\left\{\left(\hat{\sigma}_l^{-2}n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)\right\} < \varsigma_n\right)\\
&\leq \Pr\left(\left\|n^{1/2}\boldsymbol{\beta}_l + n^{1/2}\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right\|^2 < 4\sigma_l^2 B_1^{-1} m_n\varsigma_n\right)\\
&+ \Pr\left(\lambda_{\min}\left\{\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)\right\} < \frac{1}{2}B_1 m_n^{-1}\right) + \Pr\left(\hat{\sigma}_l^{-2} < \sigma_l^{-2}/2\right)\\
&\equiv A_4 + A_5 + A_6.
\end{aligned}$$

By Lemma A.5 and Lemma A.4, $A_l = \frac{1}{4}C_5\exp\left(-C_6 n^{C_7}\right)$ for some positive constants C_5, C_6, and C_7 and $l = 5, 6$. For A_4, we have with $C = 2\sigma_l B_1^{-1/2}$,

$$\begin{aligned}
A_4 &= \Pr\left(\left\|n^{1/2}\boldsymbol{\beta}_l + n^{1/2}\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right\| < C m_n^{1/2}\varsigma_n^{1/2}\right)\\
&\leq \Pr\left(\left\|n^{1/2}\boldsymbol{\beta}_l\right\| - \left\|n^{1/2}\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right\| < C m_n^{1/2}\varsigma_n^{1/2}\right)\\
&= \Pr\left(\left\|\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1}n^{-1/2}\mathbb{X}_l'\boldsymbol{u}_l\right\| > \frac{1}{2}\left\|n^{1/2}\boldsymbol{\beta}_l\right\| + \frac{1}{2}\left\|n^{1/2}\boldsymbol{\beta}_l\right\| - C m_n^{1/2}\varsigma_n^{1/2}\right)\\
&\leq \Pr\left(\left\|\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1}n^{-1/2}\mathbb{X}_l'\boldsymbol{u}_l\right\| > \frac{1}{2}\left\|n^{1/2}\boldsymbol{\beta}_l\right\|\right)\\
&\leq \Pr\left(\left\|n^{-1/2}\mathbb{X}_l'\boldsymbol{u}_l\right\| > \frac{1}{3}B_1 m_n^{-1}\left\|n^{1/2}\boldsymbol{\beta}_l\right\|\right) + \Pr\left(\lambda_{\max}\left\{\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\right\} > \frac{3}{2}B_1^{-1}m_n\right)\\
&\leq n^{-M} + \frac{1}{2}C_5\exp\left(-C_6 n^{C_7}\right)
\end{aligned}$$

for any arbitrarily large constant M and some positive constants C_5, C_6, and C_7, where the second inequality follows from the fact that $\frac{1}{2}\left\|n^{1/2}\boldsymbol{\beta}_l\right\| \gg m_n^{1/2}\varsigma_n^{1/2}$, and the last line holds by taking $v_n \propto \left(m_n^{-1/2}\left\|n^{1/2}\boldsymbol{\beta}_l\right\|\right)^2 \gtrsim \kappa_n^2 \log(m_n) m_n$ in equation (S1.14) and applying Lemma A.4. In sum, we have

$$\Pr\left(\hat{\mathcal{X}}_l < \varsigma_n\right) \leq n^{-M} + C_5\exp\left(-C_6 n^{C_7}\right),$$

or equivalently, $\Pr\left(\hat{\mathcal{X}}_l \geq \varsigma_n\right) > 1 - n^{-M} - C_5 \exp\left(-C_6 n^{C_7}\right)$. ■

Proof of Theorem 2.1. Proposition 2.1 will be used repeatedly in the proof. First, we study TPR_n. Note that

$$E\left(\mathrm{TPR}_n\right) = p^{*-1} \sum_{l=1}^{p_n} E\left[1\left(\widehat{\mathcal{J}}_l = 1 \text{ and } \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2} \neq 0\right)\right]$$
$$= p^{*-1} \sum_{l=1}^{p^*} \Pr\left(\hat{\mathcal{X}}_l \geq \varsigma_n\right) \geq 1 - n^{-M} - C_1 \exp\left(-C_2 n^{C_3}\right)$$

for some positive constants M, C_1, C_2, and C_3, where the inequality holds by the second part of Proposition 2.1 and Assumption 9.

Next, we study FPR_n.

$$E\left(\mathrm{FPR}_n\right) = \left(p_n - p^*\right)^{-1} \sum_{l=p^*+1}^{p_n} E\left[1\left(\widehat{\mathcal{J}}_l = 1 \text{ and } \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2} = 0\right)\right]$$
$$= \left(p_n - p^*\right)^{-1} \sum_{l=p^*+1}^{p^*+p^{**}} E\left[1\left(\widehat{\mathcal{J}}_l = 1 \text{ and } \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2} = 0\right)\right]$$
$$+ \left(p_n - p^*\right)^{-1} \sum_{l=p^*+p^{**}+1}^{p_n} E\left[1\left(\widehat{\mathcal{J}}_l = 1 \text{ and } \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2} = 0\right)\right]$$
$$\leq p^{**}/\left(p_n - p^*\right) + \left(p_n - p^*\right)^{-1} \sum_{l=p^*+p^{**}+1}^{p_n} E\left[1\left(\widehat{\mathcal{J}}_l = 1 \text{ and } \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2} = 0\right)\right]$$
$$\leq p^{**}/\left(p_n - p^*\right) + C_4 n^{-M} + C_5 \exp\left(-C_6 n^{C_7}\right)$$

for some positive C_4, C_5, and C_6 and any large positive constant M, and the last inequality holds by the first part of Proposition 2.1 and Assumption 9.

Now, we turn to FDR_n. Note that

$$E\left[\sum_{l=1}^{p_n} 1\left(\widehat{\mathcal{J}}_l = 1,\ \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2} = 0, \text{ and } \theta_l \lesssim \log\left(m_n\right)^{1/2}\left(m_n/n\right)^{1/2}\right)\right]$$
$$= \sum_{l=p^*+p^{**}+1}^{p_n} \Pr\left(\hat{\mathcal{X}}_l \geq \varsigma_n\right)$$
$$\leq \left(p_n - p^* - p^{**}\right)\left[n^{-M} + C_8 \exp\left(-C_9 n^{C_{10}}\right)\right]$$

for any large positive number M and some positive constants C_8, C_9, and C_{10} by the first part of Proposition 2.1. Taking a M sufficiently large, we have

$$E\left[\sum_{l=1}^{p_n} 1\left(\widehat{\mathcal{J}}_l = 1,\ \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2} = 0, \text{ and } \theta_l \lesssim \log\left(m_n\right)^{1/2}\left(m_n/n\right)^{1/2}\right)\right] \to 0.$$

Then

$$\mathrm{FDR}_n = \frac{\sum_{l=1}^{p_n} 1\left(\widehat{\mathcal{J}}_l = 1,\ \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2} = 0,\ \text{and } \theta_l \lesssim \log\left(m_n\right)^{1/2}\left(m_n/n\right)^{1/2}\right)}{\sum_{l=1}^{p_n} \widehat{\mathcal{J}}_l + 1} \xrightarrow{P} 0$$

by Markov inequality and the fact that $\sum_{l=1}^{p_n} \widehat{\mathcal{J}}_l + 1 \geq 1$. ■

Proof of Proposition 3.1. Note that $\left\|\boldsymbol{\eta}_{l,\boldsymbol{Z}}\right\| \propto m_n^{-1/2}\left\|\theta_{l,\boldsymbol{Z}}\right\|$ by equation (A.7). It is equivalent to showing the results for the case when $\boldsymbol{\eta}_{l,\boldsymbol{Z}} \lesssim \log\left(m_n\right)^{1/2} n^{-1/2}$ and the case when $\boldsymbol{\eta}_{l,\boldsymbol{Z}} \gtrsim \kappa_n\left[\log\left(m_n\right)\right]^{1/2} n^{-1/2}$.

For clarity, we begin with the case when $\boldsymbol{\eta}_{l,\boldsymbol{Z}} = 0$. In this case, the $\hat{\mathcal{X}}_{l,\boldsymbol{Z}}$ in equation (A.10) reduces to $\hat{\mathcal{X}}_{l,\boldsymbol{Z}} = \tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}'\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2 \mathbb{X}_l' M_{\mathbb{Z}} \mathbb{X}_l\right)^{-1} \tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}$, where $M_{\mathbb{Z}} = I_n - \mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'$. Then by Lemma A.12,

$$\begin{aligned}
\Pr\left(\hat{\mathcal{X}}_{l,\boldsymbol{Z}} \geq \varsigma_n\right) &\leq \exp\left(-C_1 m_n^{-1}\varsigma_n + \log m_n\right) + C_2\exp(-C_3 n^{C_4}) \\
&= \exp\left(-C_5\kappa_n \log\left(m_n\right) + \log m_n\right) + C_2\exp(-C_3 n^{C_4}) \\
&\leq n^{-M} + C_2\exp\left(-C_3 n^{C_4}\right),
\end{aligned}$$

for any arbitrarily large positive constant M, where the second line holds by the fact that $\varsigma_n \propto \kappa_n \log\left(m_n\right) m_n$ and the last line holds because $C_5\kappa_n\log\left(m_n\right) - \log m_n \gg M\log m_n$.

When $\boldsymbol{\eta}_{l,\boldsymbol{Z}} \neq 0$, we analyze $\hat{\mathcal{X}}_{l,\boldsymbol{Z}}$ similarly as we do for $\hat{\mathcal{X}}_l$ in the proof of Proposition 2.1. We first consider the case where $\boldsymbol{\eta}_{l,\boldsymbol{Z}} \lesssim \log\left(m_n\right)^{1/2} n^{-1/2}$. Note that

$$\begin{aligned}
\Pr\left(\hat{\mathcal{X}}_{l,\boldsymbol{Z}} \geq \varsigma_n\right) &= \Pr\left(\left(n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} + n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right)'\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2 n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\left(n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} + n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right) \geq \varsigma_n\right) \\
&\leq \Pr\left(\left\|n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} + n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right\|^2 \lambda_{\max}\left\{\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2 n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\} \geq \varsigma_n\right) \\
&\leq \Pr\left(\left\|n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} + n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right\|^2 \geq \frac{1}{2}\sigma^2 B_{X1} m_n^{-1}\varsigma_n\right) \\
&+ \Pr\left(\lambda_{\max}\left\{\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2 n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\} \geq 2\sigma^{-2}B_{X1}^{-1} m_n\right) \\
&\equiv A_7 + A_8. \qquad \text{(B.1)}
\end{aligned}$$

For A_7, we have

$$\begin{aligned}
A_7 &= \Pr\left(\left\|n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} + n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right\| \geq 2^{-1/2}\sigma B_{X1}^{1/2} m_n^{-1/2}\varsigma_n^{1/2}\right) \\
&\leq \Pr\left(\left\|n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}}\right\| + \left\|n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right\| \geq 2^{-1/2}\sigma B_{X1}^{1/2} m_n^{-1/2}\varsigma_n^{1/2}\right) \\
&= \Pr\left(\left\|n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right\| \geq \frac{1}{2}2^{-1/2}\sigma B_{X1}^{1/2} m_n^{-1/2}\varsigma_n^{1/2} + \frac{1}{2}2^{-1/2}\sigma B_{X1}^{1/2} m_n^{-1/2}\varsigma_n^{1/2} - \left\|n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}}\right\|\right) \\
&\leq \Pr\left(\left\|n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right\| \geq \frac{1}{2}2^{-1/2}\sigma B_{X1}^{1/2} m_n^{-1/2}\varsigma_n^{1/2}\right)
\end{aligned}$$

$$\leq n^{-M} + \frac{1}{2} C_2 \exp\left(-C_3 n^{C_4}\right), \tag{B.2}$$

where the second inequality holds by the fact that $m_n^{-1/2} \varsigma_n^{1/2} \propto \kappa_n^{1/2} [\log(m_n)]^{1/2} \gg [\log(m_n)]^{1/2} \gtrsim \left\| n^{1/2} \boldsymbol{\eta}_{l,\boldsymbol{Z}} \right\|$, and the last inequality holds by Lemma A.10. By Lemma A.9,

$$A_8 \leq \frac{1}{2} C_2 \exp\left(-C_3 n^{C_4}\right). \tag{B.3}$$

Combining (B.2), (B.3) and (B.1) yields $\Pr\left(\hat{\mathcal{X}}_{l,\boldsymbol{Z}} \geq \varsigma_n\right) \leq n^{-M} + \frac{1}{2} C_2 \exp\left(-C_3 n^{C_4}\right)$.

Now, we consider the case where $\boldsymbol{\eta}_{l,\boldsymbol{Z}} \gtrsim \kappa_n [\log(m_n)]^{1/2} n^{-1/2}$. Note that

$$\begin{aligned}
\Pr\left(\hat{\mathcal{X}}_{l,\boldsymbol{Z}} < \varsigma_n\right) &= \Pr\left(\left(n\boldsymbol{\eta}_{l,\boldsymbol{Z}} + \tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right)' \left(\hat{\sigma}_{l,\boldsymbol{Z}}^2 \mathbb{X}_l' M_{\mathbb{Z}} \mathbb{X}_l\right)^{-1} \left(n\boldsymbol{\eta}_{l,\boldsymbol{Z}} + \tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right) < \varsigma_n\right) \\
&\leq \Pr\left(\left\| n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} + n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}} \right\|^2 \lambda_{\min}\left\{\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2 n^{-1} \mathbb{X}_l' M_{\mathbb{Z}} \mathbb{X}_l\right)^{-1}\right\} < \varsigma_n\right) \\
&\leq \Pr\left(\left\| n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} + n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}} \right\|^2 < 4\sigma^2 B_{X2} m_n^{-1} \varsigma_n\right) \\
&+ \Pr\left(\lambda_{\min}\left\{\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2 n^{-1} \mathbb{X}_l' M_{\mathbb{Z}} \mathbb{X}_l\right)^{-1}\right\} < \frac{1}{4}\sigma^{-2} B_{X2}^{-1} m_n\right) \qquad \text{(B.4)} \\
&\equiv A_9 + A_{10}. \qquad \text{(B.5)}
\end{aligned}$$

Noting that $\left\| n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} + n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}} \right\| \geq \left\| n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} \right\| - \left\| n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}} \right\|$, we have

$$\begin{aligned}
A_9 &= \Pr\left(\left\| n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} + n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}} \right\| < 2\sigma B_{X2}^{1/2} m_n^{-1/2} \varsigma_n^{1/2}\right) \\
&\leq \Pr\left(\left\| n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}} \right\| > \frac{1}{2}\left\| n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} \right\| + \frac{1}{2}\left\| n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} \right\| - 2\sigma B_{X2}^{1/2} m_n^{-1/2} \varsigma_n^{1/2}\right) \\
&\leq \Pr\left(\left\| n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}} \right\| > \frac{1}{2}\left\| n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} \right\|\right) \\
&\leq \Pr\left(\left\| n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}} \right\| > C\kappa_n^{1/2} m_n^{-1/2} \varsigma_n^{1/2}\right) \\
&\leq n^{-M} + \frac{1}{2} C_5 \exp\left(-C_6 n^{C_7}\right), \qquad \text{(B.6)}
\end{aligned}$$

for any arbitrarily large positive constant M and some positive constants C_5, C_7, and C_7, where the second and third inequalities follow from the fact that $\left\| n^{1/2}\boldsymbol{\eta}_{l,\boldsymbol{Z}} \right\| \gtrsim \kappa_n [\log(m_n)]^{1/2} \propto \kappa_n^{1/2} m_n^{-1/2} \varsigma_n^{1/2} \gg 2\sigma^{-1} B_{X2}^{1/2} m_n^{-1/2} \varsigma_n^{1/2}$, and the last inequality holds by Lemma A.10 to get the last inequality. By the second part of Lemma A.9, $A_{10} \leq \frac{1}{2} C_5 \exp\left(-C_6 n^{C_7}\right)$. It follows that

$$\Pr\left(\hat{\mathcal{X}}_{l,\boldsymbol{Z}} < \varsigma_n\right) < n^{-M} + C_5 \exp\left(-C_6 n^{C_7}\right).$$

That is, $\Pr\left(\hat{\mathcal{X}}_{l,\boldsymbol{Z}} \geq \varsigma_n\right) \geq 1 - n^{-M} - C_5 \exp\left(-C_6 n^{C_7}\right)$. ■

To prove Theorem 3.1, we introduce some notations presented in Table 7 below.

Apparently, $\mathcal{D}_k = \mathcal{N}_k^c$, the complement of $\mathcal{N}_k$.

Table 7: Notations used in the proof of Theorem 3.1

Notation	Meaning
$\mathcal{B}_{l,k}$	variable l is selected at the kth stage of the OCMT procedure.
$\mathcal{L}_{l,k} = \cup_{h=1}^{k} B_{l,h}$	variable l is selected up to and including the kth stage
$\mathcal{N}_k = \cup_{l=p^*+p^{**}+1}^{p_n} L_{l,k}$	one or more noise variables are selected up to the kth stage
$\mathcal{A}_k = \cap_{l=1}^{p^*} L_{l,k}$	all signal variables are selected up to the kth stage
$\mathcal{H}_k = \cap_{l=p^*+1}^{p^*+p^{**}} L_{l,k}$	all pseudo-signal variable are selected up to the kth stage.
$\mathcal{D}_k$	variables selected up to the kth stage are signals or pseudo signals
$\mathcal{T}_k$	The OCMT procedure concludes at or before stage k

Proof of Theorem 3.1. Recall that the constants C_l's and M_l's may vary across lines. By the definition in Table 2, all pseudo-signals have strong net effects on Y in the first stage with $\theta_l \gtrsim \kappa_n \log(m_n)^{1/2}(m_n/n)^{1/2}$ for a slowly divergent series $\kappa_n \to \infty$. Under this condition, the second part of Proposition 2.1 implies all pseudo-signals can be selected in stage 1 with very high probability. That is, $\mathcal{H}_1$ happens with very high probability. Specifically,

$$\begin{aligned}\Pr(\mathcal{H}_1) &= \Pr\left(\cap_{l=p^*+1}^{p^*+p^{**}} \mathcal{B}_{l,1}\right) = 1 - \Pr\left(\cup_{l=p^*+1}^{p^*+p^{**}} \mathcal{B}_{l,1}^c\right) \geq 1 - \sum_{l=p^*+1}^{p^*+p^{**}} \Pr\left(\mathcal{B}_{l,1}^c\right) \\ &\geq 1 - p^{**}\left(n^{-M} + C_1 \exp\left(-C_2 n^{C_3}\right)\right) \\ &\geq 1 - n^{-M_1} - C_4 \exp\left(-C_5 n^{C_6}\right) \qquad \text{(B.7)}\end{aligned}$$

for some arbitrarily large constant M, some $M_1 < M - B_{p^{**}}$, and some positive constants $C_1, \ldots, C_6$, where the second inequality holds by the second part of Proposition 2.1, and the last inequality holds by Assumption 2' by restricting $p^{**} \propto n^{B_{p^{**}}}$. Since M can be arbitrarily large, M_1 can be arbitrarily large too. The above result, in conjunction with the fact that $\mathcal{H}_1 \subset \mathcal{H}_k$ for any any $k > 1$, implies that

$$\Pr(\mathcal{H}_k) \geq \Pr(\mathcal{H}_1) \geq 1 - n^{-M_1} - C_4 \exp\left(-C_5 n^{C_6}\right). \qquad \text{(B.8)}$$

That is, the probability of that all pseudo-signals are selected up to stage k is larger than the probability of that all pseudo-signals are selected up to stage 1.

Recall that p^* is the number of signals that is assumed to be fixed in Assumption 2. We consider the probability of $\mathcal{A}_{p^*} \cap \mathcal{H}_{p^*}$, the event that all signals and pseudo-signals are selected up to stage p^*. The key to bound this probability is Proposition 3.1. To apply the proposition, we need the condition that all pre-selected variables are signals or pseudo-signals, which is the event $\mathcal{D}_{p^*-1}$. In view of this observation, we conduct the analysis as follows. First,

$$\Pr(\mathcal{A}_{p^*} \cap \mathcal{H}_{p^*}) \geq \Pr(\mathcal{A}_{p^*} \cap \mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}) \qquad \text{(B.9)}$$

$$= \Pr\left(\mathcal{A}_{p^*}|\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}\right) \Pr\left(\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}\right).$$

For $\Pr\left(\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}\right)$, we have

$$\begin{aligned}
\Pr\left(\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}\right) &= 1 - \Pr\left(\mathcal{H}_{p^*}^c \cup \mathcal{D}_{p^*}^c\right) \\
&\geq \left[1 - \Pr\left(\mathcal{H}_{p^*}^c\right)\right] - \Pr\left(\mathcal{D}_{p^*-1}^c\right) \\
&\geq 1 - n^{-M_2} - C_7 \exp\left(-C_8 n^{C_9}\right) \qquad \text{(B.10)}
\end{aligned}$$

for an arbitrarily large constant M_2 and some positive constants C_7, C_8 and C_9, where the last inequality follows by equation (B.8) and Lemma A.13. Since we only have p^* signals, with Assumption 9', Lemma A.7 implies that the population effect of at least one signal on Y conditional on $\mathcal{D}_k$, $k \leq p^* - 1$, would become large enough to be picked up by our procedure with very high probability. For the lth signal, $1 \leq l \leq p^*$, we denote the stage where its effect on Y becomes large enough to be picked up by k_l^*, that is $\boldsymbol{\theta}_{l,\boldsymbol{Z}_{\left(k_l^*-1\right)}} \gtrsim \kappa_n \log\left(m_n\right)^{1/2} m_n^{1/2} n^{-1/2}$. Without loss of generality, we assume $k_1^* \leq k_2^* \leq k_3^* \ldots \leq k_{p^*}^*$. By Lemma A.7 again, $k_1^* = 1$ and $k_l^* - k_{l-1}^* \leq 1$. So the OCMT procedure with very high probability does not stop until all the signals are selected.

We now bound the probability formally. By definition,

$$\mathcal{A}_{p^*} = \cap_{l=1}^{p^*} \mathcal{L}_{l,p^*} = \cap_{l=1}^{p^*} \left(\cup_{h=1}^{p^*} \mathcal{B}_{l,h}\right) \supseteq \cap_{l=1}^{p^*} \mathcal{B}_{l,k_l^*}.$$

Noting that $k_1^* = 1$, $k_l^* - k_{l-1}^* \leq 1$, and $k_1^* \leq k_2^* \leq k_3^* \ldots \leq k_{p^*}^*$ if $\cap_{l=1}^{p^*} \mathcal{B}_{l,k_l^*}$ occurs, the OCMT procedure does not stop before stage $k_{p^*}^*$. Then,

$$\begin{aligned}
\Pr\left(\mathcal{A}_{p^*}|\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}\right) &\geq \Pr\left(\cap_{l=1}^{p^*} \mathcal{B}_{l,k_l^*}|\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}\right) \\
&= 1 - \Pr\left(\cup_{l=1}^{p^*} \mathcal{B}_{l,k_l^*}^c|\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}\right) \\
&\geq 1 - \sum_{l=1}^{p^*} \Pr\left(\mathcal{B}_{l,k_l^*}^c|\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}\right). \qquad \text{(B.11)}
\end{aligned}$$

Conditional on the event $\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}$, all variables in $\boldsymbol{Z}_{\left(k_l^*-1\right)}$ are either signals or pseudo-signals with $\boldsymbol{\theta}_{l,\boldsymbol{Z}_{\left(k_l^*-1\right)}} \gtrsim \kappa_n \log\left(m_n\right)^{1/2} m_n^{1/2} n^{-1/2}$. Then we can apply the second part of Proposition 3.1 on $\Pr\left(\mathcal{B}_{l,k_l^*}^c|\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}\right)$ to obtain

$$\Pr\left(\mathcal{B}_{l,k_l^*}^c|\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}\right) \leq n^{-M} + C_1 \exp\left(-C_2 n^{C_3}\right).$$

Substituting this into equation (B.11) yields

$$\Pr\left(\mathcal{A}_{p^*}|\mathcal{H}_{p^*} \cap \mathcal{D}_{p^*-1}\right) \geq 1 - \sum_{l=1}^{p^*} \left[n^{-M} + C_1 \exp\left(-C_2 n^{C_3}\right)\right]$$

$$\geq 1-n^{-M_3}-C_{10}\exp\left(-C_{11}n^{C_{12}}\right) \tag{B.12}$$

for some arbitrarily large constant M_3 and some positive constants C_{10}, C_{11} and C_{12}. Substituting equations (B.10) and (B.12) into equation (B.9) yields

$$\begin{aligned}\Pr\left(\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\right) &\geq \left(1-n^{-M_3}-C_{10}\exp\left(-C_{11}n^{C_{12}}\right)\right)\left(1-n^{-M_2}-C_7\exp\left(-C_8n^{C_9}\right)\right)\\ &\geq 1-n^{-M_3}-C_{10}\exp\left(-C_{11}n^{C_{12}}\right)-n^{-M_2}-C_7\exp\left(-C_8n^{C_9}\right)\\ &\geq 1-n^{-M_4}-C_{13}\exp\left(-C_{14}n^{C_{15}}\right) \end{aligned}\tag{B.13}$$

for some large positive constant M_4, and some positive constants C_{13}, C_{14}, and C_{15}.

With (B.13), we are ready to complete the proof of the first part of the theorem. Notice that

$$\begin{aligned}\Pr\left(\mathcal{T}_{p^*}\right) &\geq \Pr\left(\mathcal{T}_{p^*}\cap\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\cap\mathcal{D}_{p^*}\right)\\ &= \Pr\left(\mathcal{T}_{p^*}|\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\cap\mathcal{D}_{p^*}\right)\Pr\left(\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\cap\mathcal{D}_{p^*}\right)\\ &= \Pr\left(\mathcal{T}_{p^*}|\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\cap\mathcal{D}_{p^*}\right)\left[1-\Pr\left(\mathcal{A}^c_{p^*}\cup\mathcal{H}^c_{p^*}\cup\mathcal{D}^c_{p^*}\right)\right]\\ &\geq \Pr\left(\mathcal{T}_{p^*}|\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\cap\mathcal{D}_{p^*}\right)\left[1-\Pr\left(\mathcal{A}^c_{p^*}\cup\mathcal{H}^c_{p^*}\right)-\Pr\left(\mathcal{D}^c_{p^*}\right)\right]. \end{aligned}\tag{B.14}$$

Note that $\mathcal{T}^c_{p^*}|\left\{\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\cap\mathcal{D}_{p^*}\right\}$ is the event that the OCMT procedure does not stop after p^* stages and all signals and pseudo-signals have been selected. It is equivalent to one or more noise variables being selected at stage p^*+1 conditional on $\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\cap\mathcal{D}_{p^*}$, which is $\cup_{l=p^*+p^{**}+1}^{p_n}\mathcal{B}_{l,p^*+1}|\left\{\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\cap\mathcal{D}_{p^*}\right\}$. By the analysis in Lemma A.13,

$$\begin{aligned}\Pr\left(\mathcal{T}_{p^*}|\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\cap\mathcal{D}_{p^*}\right) &= 1-\Pr\left(\cup_{l=p^*+p^{**}+1}^{p_n}\mathcal{B}_{l,p^*+1}|\left\{\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\cap\mathcal{D}_{p^*}\right\}\right)\\ &\geq 1-\sum_{l=p^*+p^{**}+1}^{p_n}\Pr\left(\mathcal{B}_{l,p^*+1}|\left\{\mathcal{A}_{p^*}\cap\mathcal{H}_{p^*}\cap\mathcal{D}_{p^*}\right\}\right)\\ &\geq 1-p_nn^{-M}-p_nC_1\exp\left(-C_2n^{C_3}\right)\\ &\geq 1-n^{-M_5}-C_{16}\exp\left(-C_{17}n^{C_{18}}\right), \end{aligned}\tag{B.15}$$

where the last line holds for some arbitrarily large positive number M_5 and some positive constants C_{16}, C_{17} and C_{18} by the fact that $p_n\propto n^{B_p}$ in Assumption 2'. Substituting equations (B.13) and (B.15) into equation (B.14) and applying Lemma A.13 on $\Pr\left(\mathcal{D}^c_{p^*}\right)$, we obtain

$$\begin{aligned}\Pr\left(\mathcal{T}_{p^*}\right) &\geq \left[1-n^{-M_5}-C_{16}\exp(-C_{17}n^{C_{18}})\right]\\ &\quad\times\left[1-n^{-M_4}-C_{13}\exp(-C_{14}n^{C_{15}})-n^{-M}-C_1\exp(-C_2n^{C_3})\right]\\ &\geq 1-n^{-M_6}-C_{19}\exp(-C_{20}n^{C_{21}}), \end{aligned}\tag{B.16}$$

for some arbitrarily large positive number M_6 and some positive constants C_{19}, C_{20}, and C_{21}. Consequently,

$$\Pr\left(\hat{k}_s>p^*\right)=\Pr\left(\mathcal{T}^c_{p^*}\right)=1-\Pr\left(\mathcal{T}_{p^*}\right).$$

The second part of the theorem is quite straightforward given the analysis so far. First, we show the result for TPR_n. We conduct the analysis conditional on $\mathcal{D}_{p^*}$ that all selected variables up to stage p^* are either signals or pseudo-signals. In the first stage, by Lemma A.7, there exists at least one signal with $\theta_l \gtrsim \kappa_n \log(m_n)^{1/2} m_n^{1/2} n^{-1/2}$. Let $\Psi^0_{(1)} = \left\{l : 1 \le l \le p^* \text{ and } \theta_l \gtrsim \kappa_n \log(m_n)^{1/2} m_n^{1/2} n^{-1/2}\right\}$, the collection of signals that have large effects on Y in the first stage. By the same logic as we have used for the proof to the first part of the theorem,

$$\begin{aligned}\Pr\left(\cap_{l\in\Psi^0_{(1)}}\left\{\widehat{\mathcal{J}}_{l,(1)}=1\right\}\right) &= 1-\Pr\left(\cup_{l\in\Psi^0_{(1)}}\left\{\widehat{\mathcal{J}}_{l,(1)}=0\right\}\right)\\ &\ge 1-n^{-M}-C_1\exp\left(-C_2 n^{C_3}\right)\end{aligned}$$

for some large positive constant M and some positive constants C_1, C_2 and C_3, where we use the fact that p^* is a fixed number and the number of elements in $\Psi^0_{(1)}$ is p^* at most. Conditional on $\mathcal{D}_{p^*}$, denote the pre-selected variables as $\boldsymbol{Z}_1$ and the index set of $\boldsymbol{Z}_1$ as $S_{(1)}$ for stage 2. Note that there may be some hidden signals accidentally selected in stage 1. We include all those in $\boldsymbol{Z}_1$ too. As long as we conduct the analysis conditional on $\mathcal{D}_{p^*}$ that no noise variables are selected, we can proceed the analysis as usual. Let $\Psi^0_{(2)} = \left\{l : 1 \le l \le p^*,\ l \notin S_{(1)} \text{ and } \theta_{l,\boldsymbol{Z}_1} \gtrsim \kappa_n \log(m_n)^{1/2} m_n^{1/2} n^{-1/2}\right\}$, the set of signals that have large effects on Y with pre-selected variable $\boldsymbol{Z}_1$. By Lemma A.7, $\Psi^0_{(2)}$ is not empty as long as $S_{(1)} \ne \{1, 2, \ldots, p^*\}$ that there are some hidden signals. For the same reason,

$$\Pr\left(\cap_{l\in\Psi^0_{(2)}}\left\{\widehat{\mathcal{J}}_{l,(2)}=1\right\}\middle|\cap_{l\in\Psi^0_{(1)}}\left\{\widehat{\mathcal{J}}_{l,(1)}=1\right\},\mathcal{D}_{p^*}\right)\ge 1-n^{-M}-C_1\exp\left(-C_2 n^{C_3}\right).$$

So on and so forth. Suppose there are some remaining signals as k_s^* in the last stage. Because we only have p^* signals, $k_s^* \le p^*$. We similarly use $\Psi^0_{(k)}$ denote set of the signals that have large effects on Y with pre-selected variables $\boldsymbol{Z}_{k-1}$ and conditional on $\mathcal{D}_{p^*}$. Then

$$\Pr\left(\cap_{l\in\Psi^0_{(k_s^*)}}\left\{\widehat{\mathcal{J}}_{l,(k_s^*)}=1\right\}\middle|\cap_{k=1}^{k_s^*-1}\cap_{l\in\Psi^0_{(k)}}\left\{\widehat{\mathcal{J}}_{l,(k)}=1\right\},\mathcal{D}_{p^*}\right)\ge 1-n^{-M}-C_1\exp\left(-C_2 n^{C_3}\right).$$

Further,

$$\begin{aligned}\Pr\left(\cap_{k=1}^{k_s^*}\cap_{l\in\Psi^0_{(k)}}\left\{\widehat{\mathcal{J}}_{l,(k)}=1\right\}\right) &\ge \Pr\left(\cap_{k=1}^{k_s^*}\cap_{l\in\Psi^0_{(k)}}\left\{\widehat{\mathcal{J}}_{l,(k)}=1\right\}\cap\mathcal{D}_{p^*}\right)\\ &= \Pr\left(\cap_{k=1}^{k_s^*}\cap_{l\in\Psi^0_{(k)}}\left\{\widehat{\mathcal{J}}_{l,(k)}=1\right\}\middle|\mathcal{D}_{p^*}\right)\Pr\left(\mathcal{D}_{p^*}\right)\\ &= \Pr\left(\cap_{l\in\Psi^0_{(k_s^*)}}\left\{\widehat{\mathcal{J}}_{l,(k_s^*)}=1\right\}\middle|\cap_{k=1}^{k_s^*-1}\cap_{l\in\Psi^0_{(k)}}\left\{\widehat{\mathcal{J}}_{l,(k)}=1\right\},\mathcal{D}_{p^*}\right)\\ &\quad\times\Pr\left(\cap_{l\in\Psi^0_{(k_s^*-1)}}\left\{\widehat{\mathcal{J}}_{l,(k_s^*-1)}=1\right\}\middle|\cap_{k=1}^{k_s^*-2}\cap_{l\in\Psi^0_{(k)}}\left\{\widehat{\mathcal{J}}_{l,(k)}=1\right\},\mathcal{D}_{p^*}\right)\\ &\quad\times\cdots\\ &\quad\times\Pr\left(\cap_{l\in\Psi^0_{(1)}}\left\{\widehat{\mathcal{J}}_{l,(1)}=1\right\}\middle|\mathcal{D}_{p^*}\right)\Pr\left(\mathcal{D}_{p^*}\right)\\ &\ge\left[1-n^{-M}-C_1\exp\left(-C_2 n^{C_3}\right)\right]^{k_s^*+1}\end{aligned}$$

$$
\begin{aligned}
&\geq 1-\left(k_s^*+1\right) n^{-M}-\left(k_s^*+1\right) C_1 \exp \left(-C_2 n^{C_3}\right) \\
&\geq 1-n^{-M_7}-C_{22} \exp \left(-C_{23} n^{C_{24}}\right)
\end{aligned}
$$

for some large positive constant M_7 and some positive constants C_{22}, C_{23}, and C_{24}, where we use the inequalities developed right before this equation and Lemma A.13 to obtain the second inequality. Consequently, we have

$$
\begin{aligned}
E\left(\mathrm{TPR}_n\right) &= p^{*-1} \sum_{l=1}^{p_n} E\left\{1\left(\widehat{\mathcal{J}}_l=1 \text{ and } \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2} \neq 0\right)\right\} \\
&= p^{*-1} \sum_{l=1}^{p^*} \operatorname{Pr}\left(\widehat{\mathcal{J}}_l=1\right) \\
&\geq p^{*-1} \sum_{l=1}^{p^*} \operatorname{Pr}\left(\left\{\widehat{\mathcal{J}}_l=1\right\} \cap\left\{\cap_{k=1}^{k_s^*} \cap_{l \in \Psi_{(k)}^0}\left\{\widehat{\mathcal{J}}_{l,(k)}=1\right\}\right\}\right) \\
&= p^{*-1} \sum_{l=1}^{p^*} \operatorname{Pr}\left(\cap_{k=1}^{k_s^*} \cap_{l \in \Psi_{(k)}^0}\left\{\widehat{\mathcal{J}}_{l,(k)}=1\right\}\right) \\
&\geq 1-n^{-M_7}-C_{22} \exp \left(-C_{23} n^{C_{24}}\right),
\end{aligned}
$$

where the fourth line follows from the fact that

$$
\left\{\widehat{\mathcal{J}}_l=1\right\} \cap\left\{\cap_{k=1}^{k_s^*} \cap_{l \in \Psi_{(k)}^0}\left\{\widehat{\mathcal{J}}_{l,(k)}=1\right\}\right\}=\left\{\cap_{k=1}^{k_s^*} \cap_{l \in \Psi_{(k)}^0}\left\{\widehat{\mathcal{J}}_{l,(k)}=1\right\}\right\}
$$

for $1 \leq l \leq p^*$. The above equality holds by the way we define k_s^* to ensure the effect of each signal on Y becomes large enough at a certain stage.

Next, we turn to FPR_n.

$$
\begin{aligned}
E\left(\mathrm{FPR}_n\right) &= \left(p_n-p^*\right)^{-1} \sum_{l=p^*+1}^{p_n} E\left\{1\left(\widehat{\mathcal{J}}_l=1 \text{ and } \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2}=0\right)\right\} \\
&= \left(p_n-p^*\right)^{-1} \sum_{l=p^*+1}^{p^*+p^{**}} E\left[1\left(\widehat{\mathcal{J}}_l=1\right)\right]+\left(p_n-p^*\right)^{-1} \sum_{l=p^*+p^{**}+1}^{n} \operatorname{Pr}\left(\widehat{\mathcal{J}}_l=1\right) \\
&\leq \frac{p^{**}}{p_n-p^*}+\left(p_n-p^*\right)^{-1} \sum_{l=p^*+p^{**}+1}^{p_n}\left[\operatorname{Pr}\left(\left(\widehat{\mathcal{J}}_l=1\right) \cap \mathcal{T}_{p^*}\right)+\operatorname{Pr}\left(\left(\widehat{\mathcal{J}}_l=1\right) \cap \mathcal{T}_{p^*}^c\right)\right] \\
&\leq \frac{p^{**}}{p_n-p^*}+\left(p_n-p^*\right)^{-1} \sum_{l=p^*+p^{**}+1}^{p_n}\left[\operatorname{Pr}\left(\left(\widehat{\mathcal{J}}_l=1\right) \middle| \mathcal{T}_{p^*}\right) \operatorname{Pr}\left(\mathcal{T}_{p^*}\right)+\operatorname{Pr}\left(\mathcal{T}_{p^*}^c\right)\right] \\
&\leq \frac{p^{**}}{p_n-p^*}+\operatorname{Pr}\left(\mathcal{N}_{p^*} \mid \mathcal{T}_{p^*}\right) \operatorname{Pr}\left(\mathcal{T}_{p^*}\right)+\operatorname{Pr}\left(\mathcal{T}_{p^*}^c\right) \\
&= \frac{p^{**}}{p_n-p^*}+\left[1-\operatorname{Pr}\left(\mathcal{D}_{p^*} \mid \mathcal{T}_{p^*}\right)\right] \operatorname{Pr}\left(\mathcal{T}_{p^*}\right)+\operatorname{Pr}\left(\mathcal{T}_{p^*}^c\right) \\
&\leq \frac{p^{**}}{p_n-p^*}+n^{-M_7}+C_{22} \exp \left(-C_{23} n^{C_{24}}\right),
\end{aligned}
$$

for some large positive constant M_7 and positive constants C_{22}, C_{23}, and C_{24}, where the fifth line holds by the fact that conditioning on $\mathcal{T}_{p^*}$, $\cup_{l=p^*+p^{**}+1}^{p_n}\left\{\widehat{\mathcal{J}}_l=1\right\} \subseteq \mathcal{N}_{p^*}$ (conditional on the OCMT procedure stops before stage p^*, whether one or more noise variables selected by OCMT is a subset of whether one or more noise variables selected before stage p^*), and the last inequality holds by applying Lemma A.13 and equation (B.15).

Now, we turn to FDR_n. By the same analysis for FPR_n, we have

$$\begin{aligned}
& E\left[\sum_{l=1}^{p_n} 1\left(\widehat{\mathcal{J}}_l=1,\ \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2}=0,\ \text{and } \theta_l \lesssim \log\left(m_n\right)^{1/2}\left(m_n/n\right)^{1/2}\right)\right] \\
& = \sum_{l=p^*+p^{**}+1}^{p_n} E\left(\widehat{\mathcal{J}}_l=1\right) \\
& \leq \sum_{l=p^*+p^{**}+1}^{p_n}\left[\Pr\left(\left(\widehat{\mathcal{J}}_l=1\right) \cap \mathcal{T}_{p^*}\right)+\Pr\left(\left(\widehat{\mathcal{J}}_l=1\right) \cap \mathcal{T}_{p^*}^c\right)\right] \\
& \leq n^{-M_7}+C_{22} \exp\left(-C_{23} n^{C_{24}}\right)=o(1).
\end{aligned}$$

Then

$$\text{FDR}_n=\frac{\sum_{l=1}^{p_n} 1\left(\widehat{\mathcal{J}}_l=1,\ \left\{E\left[f_l^*\left(X_l\right)^2\right]\right\}^{1/2}=0,\ \text{and } \theta_l \lesssim \log\left(m_n\right)^{1/2}\left(m_n/n\right)^{1/2}\right)}{\sum_{l=1}^{p_n} \widehat{\mathcal{J}}_l+1} \xrightarrow{P} 0.$$

by Markov inequality and the fact that $\sum_{l=1}^{p_n} \widehat{\mathcal{J}}_l+1 \geq 1$. ■

Proof of Theorem 3.2. For (i), we denote the event of the desirable results from the OCMT procedure as

$$\mathcal{M}_{\text{OCMT}}=\mathcal{T}_{p^*} \cap \mathcal{A}_{p^*} \cap \mathcal{H}_{p^*} \cap \mathcal{D}_{p^*},$$

which is the event that the OCMT concludes at or before stage p^*, all the signals and pseudo signals are selected, and none of the noise variables are included. By the proof of Theorem 3.1 (see esp. equations (B.13) and (B.15)) and by applying Lemma A.13 on $\Pr\left(\mathcal{D}_{p^*}^c\right)$, we have

$$\Pr\left(\mathcal{M}_{\text{OCMT}}\right)=1-o\left(1\right). \tag{B.17}$$

Conditional on $\mathcal{M}_{\text{OCMT}}$, we claim that all the technical conditions in Theorem 3 of Huang, Horowitz and Wei (2010) (HHW) are satisfied. For a better exposition, we defer the proof of this claim to the end of the proof.

Now, we denote the desirable event of the adaptive group Lasso as

$$\mathcal{M}_{\text{AGLASSO}}=\{\text{All signals are selected, no pseudo-signals or noise variables are selected}\}.$$

Theorem 3 of HHW implies that

$$\Pr\left(\mathcal{M}_{\text{AGLASSO}} \mid \mathcal{M}_{\text{OCMT}}\right)=1-o\left(1\right). \tag{B.18}$$

Consequently we have

$$\begin{aligned}\Pr\left(\mathcal{M}_{\text{AGLASSO}}\right) &= \Pr\left(\mathcal{M}_{\text{AGLASSO}}|\,\mathcal{M}_{\text{OCMT}}\right)\Pr\left(\mathcal{M}_{\text{OCMT}}\right)+\Pr\left(\mathcal{M}_{\text{AGLASSO}}|\,\mathcal{M}^{c}_{\text{OCMT}}\right)\Pr\left(\mathcal{M}^{c}_{\text{OCMT}}\right)\\ &\geq \Pr\left(\mathcal{M}_{\text{AGLASSO}}|\,\mathcal{M}_{\text{OCMT}}\right)\Pr\left(\mathcal{M}_{\text{OCMT}}\right)+o(1)\\ &= 1-o\left(1\right),\end{aligned}$$

where we use the results in equations (B.17) and (B.18).

We turn to (ii). By the result in (i), $\Pr\left(\mathcal{M}_{\text{AGLASSO}}\right) > 1-\varepsilon/2$ for any fixed small positive value ε after some large n. Conditional on $\mathcal{M}_{\text{AGLASSO}}$, the post OCMT estimation is simply a special case of Stone (1985) because p^* is fixed. Further, the bias term is of the order m_n^{-d}. Since $m_n^{-d} \ll (m_n/n)^{1/2}$ by Assumption 7. That is, the bias term is asymptotically negligible. Then the results in Stone (1985), we have conditional on $\mathcal{M}_{\text{AGLASSO}}$,

$$P^{m_n}\left(\boldsymbol{Z}_{\text{AGLASSO}}\right)'\hat{\boldsymbol{\beta}}_{\text{post}}-\sum_{j=1}^{p^*}f_j^*\left(X_j\right)=O_P\left(\left(m_n/n\right)^{1/2}\right). \tag{B.19}$$

Then unconditionally, we have $P^{m_n}\left(\boldsymbol{Z}_{\text{AGLASSO}}\right)'\hat{\boldsymbol{\beta}}_{\text{post}}-\sum_{j=1}^{p^*}f_j^*\left(X_j\right)=O_P\left(\left(m_n/n\right)^{1/2}\right)$ as we can make ε arbitrarily small.[13].

We complete the proof by demonstrating the initial claim. To achieve this, we need to verify the conditions for the tuning parameters and conditions A1–A4 as specified in HHW.

- First, conditional on $\mathcal{M}_{\text{OCMT}}$, all signals and pseudo signals are selected, while none of the noise variables are included. Consequently, the regular rank condition, as imposed in Assumption 10, holds. Furthermore, the number of covariates is p^*+p^{**}, which is of the same order as p^{**} when $p^{**}>0$ because p^* is fixed.

- Second, we discuss the tuning parameters. The condition $\lambda_{n1}\geq C\sqrt{n\log(p^{**}m_n)}$ was directly imposed in Theorem 1 of HHW. The condition $\lambda_{n1}\ll\sqrt{n/m_n}$ is stronger than the one imposed in part (ii) of Theorem 1 of HHW. Thus, λ_{n1} satisfies the requirements in HHW. For λ_{n2}, we only need to verify whether it satisfies condition (B2) in HHW. The condition $\lambda_{n2}\ll nm_n^{-1/4}$ clearly satisfies B2(a) in HHW. For B2(b) in HHW, firstly, $r_n \propto \sqrt{n/[m_n\log(p^{**}m_n)]}$ in our case (the convergence rate ensured by Theorem 1 of HHW), because $\lambda_{n1}\ll\sqrt{n/m_n}$ and the bias term $m_n^{-d}\ll\sqrt{m_n/n}$ as ensured by Assumption 7. Using the rate of r_n, some simple

[13] This holds because for any events A and B with $\Pr\left(B\right)>1-\varepsilon/2$ and $\Pr\left(A|B\right)>1-\varepsilon/2$, we have

$$\Pr\left(A\right)=\Pr\left(A|B\right)\Pr\left(B\right)+\Pr\left(A|B^c\right)\Pr\left(B^c\right)\geq\Pr\left(A|B\right)\Pr\left(B\right)>1-\varepsilon.$$

calculations show that $\lambda_{n2} \gg m_n^{1/2} \log(p^{**} m_n)$ satisfies B2(b), again using Assumption 7. We have verified the conditions required for the tuning parameters.

- Third, condition A3 in HHW is merely a normalization, and we conduct such a normalization as well. The major parts of conditions A1, A2, and A4 in HHW have been imposed in this paper except for two conditions:

  1. HHW impose $\min_{1 \le j \le p^*} \left\| f_j^* \right\| \ge C > 0$, while our counterpart is $\left\{E[f_j^*(X_j)^2]\right\}^{1/2} \gtrsim \kappa_n \log(m_n)^{1/2}(m_n/n)^{1/2}$ for $j = 1, 2, ..., p^*$ as in Assumption 9'.
  2. HHW assume $\Pr(|\varepsilon_i| > t) \le K \exp(-Ct^2)$ for some positive constants K and C and for all t, while we assume $\Pr(|\varepsilon_i| > t) \le K \exp(-Ct^s)$ for some constant $s > 0$ in Assumption 4. Clearly, our conditions are weaker.

  We are going to show that these two conditions have no impact on the final results.

For 1. As stated in the second bullet point, the rate of convergence of the Lasso estimator is $r_n^{-1} \propto \sqrt{m_n \log(p^{**} m_n)/n}$. Note that $\log(p^{**} m_n) \propto \log(m_n)$ under our framework, because $m_n \propto n^{B_m}$ and $p^{**} \propto n^{B_{p^{**}}}$. Therefore, $\{E[f_j^*(X_j)^2]\}^{1/2} \gg r_n^{-1}$ for $j = 1, 2, ..., p^*$, due to the fact that $\kappa_n \to \infty$. This result implies that the Lasso estimator of signals will not be zero with very high probability because the signal strength is much greater than the convergence rate. Therefore, this weaker condition has no impact on the Lasso shrinkage result.

For 2. Our tail condition slightly generalizes the one in HHW. The major use of this condition in HHW is the probability bounds derived in Lemma 2 of HHW. In particular, only the second part of this lemma is useful for us, viz., the result when $\frac{m_n \log(p m_m)}{n} \to 0$, the case in this paper. To guarantee this result for our case, we can apply Lemma A.2 in this paper by setting $v_n = C n^{1/2} m_m^{-1/2} \sqrt{\log(p_n m_n)}$ for a sufficiently large constant C on the series $\sum_{i=1}^n \phi_j(x_{ki})\varepsilon_i$ for $j = 1, ..., m_n$ and $k = 1, ..., p_n$. The probability bound from this lemma is $\exp\left[-\frac{C^2(1-\pi)^2 \log(p_n m_n)}{2C_1}\right]$ for another uniformly bounded C_1, for $j = 1, ..., m_n$ and $k = 1, ..., p_n$. We set C large enough so that the probability bound is smaller than $(p_n m_n)^{-M}$ for some $M > 1$. Then

$$\Pr\left(\max_{j=1,...,m_n,k=1,...,p_n} \left|\sum_{i=1}^n \phi_j(x_{ki})\varepsilon_i\right| > v_n\right) \le \sum_{j=1}^{m_n}\sum_{k=1}^{p_n} \Pr\left(\left|\sum_{i=1}^n \phi_j(x_{ki})\varepsilon_i\right| > v_n\right)$$
$$\le m_n p_n (p_n m_n)^{-M} \to 0,$$

which implies $\max_{j=1,...,m_n,k=1,...,p_n} |\sum_{i=1}^n \phi_j(x_{ki})\varepsilon_i| = O_P\left[n^{1/2} m_m^{-1/2} \sqrt{\log(p_n m_n)}\right]$, an equivalent result to the second part of Lemma 2 in HHW.

■

Proof of Theorem 3.3. The uniform boundedness of $\sum_{j=1}^{p^*} f_j(X_j)$ implies that U_l also satisfies the tail condition in Assumption 4. [For details, see the proof of Lemma A.5.] This, in conjunction with the additional conditions in Assumption 2", implies that all technical lemmas can go through. Consequently, we only need to show that the probability bounds for the events like $\mathcal{T}_{p^*}, \mathcal{A}_{p^*}, \mathcal{H}_{p^*}$, and $\mathcal{D}_{p^*}$ are still $1 - o(1)$ with a diverging p^*. This statement holds due to the fact that the error bounds obtained before are of the order either n^{-M} or $\exp\left(-n^C\right)$ for some large positive constant M and some $C > 0$, and the error probabilities accumulated for a diverging p^* are of the order $p^*\left(n^{-M} + \exp\left(-n^C\right)\right) = o(1)$. We provide some details below.

We continue to use M to denote some large positive constant and C to denote some generic positive constant. For $\mathcal{H}_{p^*}$, equations (B.7) and (B.8) imply that

$$\Pr\left(\mathcal{H}_{p^*}\right) \geq \Pr\left(\mathcal{H}_1\right) = \Pr\left(\cap_{l=p^*+1}^{p^*+p^{**}} \mathcal{B}_{l,1}\right) = 1 - \Pr\left(\cup_{l=p^*+1}^{p^*+p^{**}} \mathcal{B}_{l,1}^c\right) \geq 1 - \sum_{l=p^*+1}^{p^*+p^{**}} \Pr\left(\mathcal{B}_{l,1}^c\right)$$
$$\geq 1 - p^{**}\left(n^{-M} + C_1 \exp\left(-C_2 n^{C_3}\right)\right) \geq 1 - n^{-M_1} - C_4 \exp\left(-C_5 n^{C_6}\right).$$

For $\mathcal{D}_{p^*}$,

$$\Pr\left(\mathcal{D}_{p^*}\right) \geq 1 - n^{-M_2} - C_7 \exp\left(-C_8 n^{C_9}\right)$$

holds by Lemma A.13 because $p^* \lesssim n^{B_{p^*}}$. Similarly,

$$\Pr\left(\mathcal{A}_{p^*} \cap \mathcal{H}_{p^*}\right) \geq 1 - n^{-M_3} - C_{10} \exp\left(-C_{11} n^{C_{12}}\right)$$

for the same reason as we obtain equation (B.13) and $p^* \lesssim n^{B_{p^*}}$. The probability bound for $\mathcal{T}_{p^*}$ was derived based on the bounds for $\mathcal{A}_{p^*}, \mathcal{H}_{p^*}$, and $\mathcal{D}_{p^*}$. Since the probability bounds for $\mathcal{A}_{p^*}, \mathcal{H}_{p^*}$, and $\mathcal{D}_{p^*}$ have been shown, the probability bound for $\mathcal{T}_{p^*}$ can be obtained using the same idea as we get equation (B.16). That is,

$$\Pr\left(\mathcal{T}_{p^*}\right) \geq 1 - n^{-M_4} - C_{13} \exp\left(-C_{14} n^{C_{15}}\right).$$

Then the first result is reached by observing that $\Pr\left(\hat{k}_s > p^*\right) = 1 - \Pr\left(\mathcal{T}_{p^*}\right)$. The results of TPR, FPR, and FDR can be proved similarly as in the proof of Theorem 3.1 and thus omitted. ■

References

Ahmed, R., and Pesaran M. H. (2022). "Regional Heterogeneity and U.S. Presidential Elections: Real-Time 2020 Forecasts and Evaluation," *International Journal of Forecasting,* 38, 662–687.

Akaike, H. (1973). "Information Theory and an Extension of Maximum Likelihood Principle," in *Proceedings of Second International Symposium on Information Theory,* Petrov, B. N. and Csaki, F. (eds.), pp. 267-281.

Akaike, H. (1974). “A New Look at the Statistical Model Identification,” *IEEE Transactions on Automatic Control,* 19(6), 716–723.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012): “Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain,” *Econometrica*, 80(6), 2369–2429.

Belloni, A., and Chernozhukov, V. (2013): “Least Squares after Model Selection in High-dimensional Sparse Models,” *Bernoulli*, 19(2), 521–547.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014): “Inference on Treatment Effects after Selection among High-Dimensional Controls,” *The Review of Economic Studies*, 81(2), 608–650.

Belloni, A., Chernozhukov, V., Fernandez-val, I., and Hansen, C. (2017): “Program Evaluation and Causal Inference with High-Dimensional Data,” *Econometrica*, 85(1), 233–298.

Bickel P. J., Ritov, Y., and Tsybakov, A. B. (2009): “Simultaneous Analysis of Lasso and Dantzig Selector,” *The Annals of Statistics*, 37(4), 1705–1732.

Buhlmann, p., and van de geer, S. (2011): *Statistics for High Dimensional Data: Methods, Theory and Applications.* Springer, Heidelberg.

Cai, Q. (2003): “Migrant Remittances and Family Ties: a Case Study in China,” *International Journal of Population Geography*, 9(6), 471–483.

Chen, X. (2007): “Large Sample Sieve Estimation of Semi-nonparametric Models,” *Handbook of Econometrics*, Chapter 76, Vol. 6B, North-Holland.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018): “Double/debiased Machine Learning for Treatment and Structural Parameters,” *The Econometric Journal,* 21(1), C1–C68.

Chudik, A., Kapetanios, G., and Pesaran, M. H. (2018): “A One Covariate at a Time, Multiple Testing Approach to Variable Selection in High-Dimensional Linear Regression Models,” *Econometrica*, 86(4), 1479–1512.

Chudik, A., M., Pesaran, M. H., and Sharifvaghefi M. (2021): “Variable Selection and Forecasting in High Dimensional Linear Regressions with Parameter Instability,” *Working Paper.*

de Boor, C. (2001): *A Practical Guide to Splines*, Springer-Verlag, New York.

Fan, J. and Li, R. (2001): “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties,” *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J., Li, R., Zhang, C. H., and Zou, H. (2020): *Statistical Foundations of Data Science*, Chapman and Hall/CRC.

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society, Series B*, 70, 849–911.

Fan, J., and Lv, J. (2013): "Asymptotic Equivalence of Regularization Methods in Thresholded Parameter Space," *Journal of the American Statistical Association*, 108, 1044–1061.

Fan, J., Feng, Y., Song, R. (2011): "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106(494), 544-557.

Fan J., Samworth, R., and Wu, Y. (2009): "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 2013–2038.

Fan, J., and R. Song (2010): "Sure Independence Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics,* 38, 3567–3604.

Fan, J., and Tang, C.(2013): "Tuning Parameter Selection in High Dimensional Penalized Likelihood," *Journal of the Royal Statistical Society, Series B*, 75, 531–552.

Gong, X., Kong, S. T., Li, S., and Meng X. (2008): "Rural-urban Migrants: a Driving Force for Growth," in Song, L. and Woo, W. T. (eds.) *China's Dilemma*, Canberra: Asia Pacific Press.

Hansen, B. (2014): "Nonparametric Sieve Regression: Least Squares, Averaging Least Squares, and Cross-Validation," Chapter 8 in *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, edited by Racine, J., Su, L., and Ullah, A..

Horowitz, J. L., and Mammen, E. (2004): "Nonparametric Estimation of an Additive Model with a Link Function," *The Annals of Statistics*, 32(6), 2412–2443.

Huang, J., Horowitz, J. L., and Wei, F. (2010): "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38(4), 2282–2313.

Kozbur, d. (2020): "Analysis of Testing-Based Forward Model Selection," *Econometrica*, 38(4), 2282–2313.

Li, Q. (2000): "Efficient Estimation of Additive Partially Linear Models," *International Economic Review*, 41(4), 1073–1092.

Li, Q. (2001): "Analysis of the Remittance of the Out-Migrant Workers in China," *Sociological Research*, 4, 64–76.

NEWEY, W. K., (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79(1), 147–168.

RACINE, J., (2022): "A Primer on Regression Splines," *Working Paper*, Available at https://cran.ma.ic.ac.uk/web/packages/crs/vignettes/spline_primer.pdf.

ROZELLE, S., TAYLOR, J. E, AND DEBRAUW, A. (1999): "Migration, Remittances, and Agricultural Productivity in China," *The American Economic Review*, 89(2), 287–291.

SCHWARZ, G. (1978) "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 471.

SHARIFVAGHEFI, M. (2023): "Variable Selection in Linear Regressions with Many Highly Correlated Covariates." *Working Paper*, Available at SSRN: http://dx.doi.org/10.2139/ssrn.4159979.

STONE, C. J. (1985): "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13(2), 689–705.

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288

VAN DER VAART, A. W., AND WELLNER, J. A. (1996): *Weak Convergence and Empirical Processes*, New York, Springer.

WU, T. T., AND LANGE, K. (1998): "Coordinate Descent Algorithms for Lasso Penalized Regression," *The Annals of Statistics*, 2(1), 224–244.

ZOU, H., AND HASTIE, T. (2005): "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320.

ZOU, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association,* 101, 1418–1429.

ZOU, H., AND LI, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics*, 36, 1509–1533.

ZHANG, C. H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942.

ZHOU, S., SHEN, X., AND WOLFE, D. A., (1998): "Local Asymptotics for Regression Splines and Confidence Regions," *The Annals of Statistics*, 26(5), 1760–1782.

Online Appendix to

"A One Covariate at a Time Multiple Testing Approach to Variable Selection in Additive Models"

(NOT for Publication)

Liangjun Su[1], Thomas Tao Yang[2], Yonghui Zhang[3], and Qiankun Zhou[4]

This supplement is composed of two parts. Section S1 contains the proofs of Lemmas A.4-A.12. Section S2 provides some additional simulation and application results.

S1 Proofs of the Technical Lemmas

Proof of Lemma A.4. To prove the lemma, we first present two inequalities: for any $m_n \times m_n$ symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we have

$$\max\left\{\left|\lambda_{\min}\left(\boldsymbol{A}\right)\right|, \left|\lambda_{\max}\left(\boldsymbol{A}\right)\right|\right\} \le m_n \left\|\boldsymbol{A}\right\|_{\infty}, \tag{S1.1}$$

and

$$\left|\lambda_{\min}\left(\boldsymbol{A}\right) - \lambda_{\min}\left(\boldsymbol{B}\right)\right| \le \max\left\{\left|\lambda_{\min}\left(\boldsymbol{A}-\boldsymbol{B}\right)\right|, \left|\lambda_{\min}\left(\boldsymbol{B}-\boldsymbol{A}\right)\right|\right\}. \tag{S1.2}$$

We will prove them at the end of this proof.

Recall that $\Phi_{X_l} = E\left[P^{m_n}\left(X_l\right)P^{m_n}\left(X_l\right)'\right]$. Let $\Phi_{n,l} = n^{-1}\mathbb{X}_l'\mathbb{X}_l$. The (j,k)-th entry of $\Phi_{n,l} - \Phi_l$ is $\xi_{jk,l} \equiv n^{-1}\sum_{i=1}^{n}\phi_j\left(x_{li}\right)\phi_k\left(x_{li}\right) - E\left[\phi_j\left(X_l\right)\phi_k\left(X_l\right)\right]$. Notice that

$$\operatorname{var}\left(\phi_j\left(X_l\right)\phi_k\left(X_l\right)\right) \le E\left[\phi_j\left(X_l\right)^2\phi_k\left(X_l\right)^2\right] \le E\left[\phi_j\left(X_l\right)^2\right] \le B_4 m_n^{-1},$$

where the second inequality holds by $\|\phi_k\|_{\infty} \le 1$, and the last inequality holds by Lemma A.3. By Lemma A.1,

$$\Pr\left(\left|\xi_{jk,l}\right| \ge v_n/n\right) \le 2\exp\left\{-v_n^2/\left[2\left(B_4 n m_n^{-1} + v_n/3\right)\right]\right\}.$$

Then by the union bound,

$$\begin{aligned}\Pr\left(\left\|\Phi_{n,l} - \Phi_{X_l}\right\|_{\infty} \ge v_n/n\right) &= \Pr\left(\max_{j,k=1,2,\ldots,m_n}\left|\xi_{jk,l}\right| \ge v_n/n\right) \\ &\le \sum_{j=1}^{m_n}\sum_{k=1}^{m_n}\Pr\left(\left|\xi_{jk,l}\right| \ge v_n/n\right) \\ &\le 2m_n^2\exp\left\{-v_n^2/[2\left(B_4 n m_n^{-1} + v_n/3\right)]\right\}. \end{aligned} \tag{S1.3}$$

[1] School of Economics and Management, Tsinghua University

[2] Research School of Economics, Australian National University

[3] School of Economics, Renmin University of China

[4] Department of Economics, Louisiana State University

Note that $\left\|\Phi_{n,l}^{-1}\right\| - \left\|\Phi_{X_l}^{-1}\right\| = \left[\lambda_{\min}\left(\Phi_{n,l}\right)\right]^{-1} - \left[\lambda_{\min}\left(\Phi_{X_l}\right)\right]^{-1}$. By (S1.1), (S1.2), and (S1.3),

$$\begin{aligned}
&\Pr\left(\left|\lambda_{\min}\left(\Phi_{n,l}\right) - \lambda_{\min}\left(\Phi_{X_l}\right)\right| \geq m_n v_n/n\right) \\
&\leq \Pr\left(\max\left\{\left|\lambda_{\min}\left(\Phi_{n,l} - \Phi_{X_l}\right)\right|, \left|\lambda_{\min}\left(\Phi_{X_l} - \Phi_{n,l}\right)\right|\right\} \geq m_n v_n/n\right) \\
&\leq \Pr\left(\left\|\Phi_{n,l} - \Phi_{X_l}\right\|_\infty \geq v_n/n\right) \\
&\leq 2m_n^2 \exp\left\{-v_n^2/\left[2\left(B_4 n m_n^{-1} + v_n/3\right)\right]\right\}. \qquad \text{(S1.4)}
\end{aligned}$$

Set $v_n = n m_n^{-2} B_1/3$. Then the above inequality becomes

$$\Pr\left(\left|\lambda_{\min}\left(\Phi_{n,l}\right) - \lambda_{\min}\left(\Phi_{X_l}\right)\right| \geq B_1 m_n^{-1}/3\right) \leq 2m_n^2 \exp\left\{-C_1 n m_n^{-3}\right\} \qquad \text{(S1.5)}$$

for some constant C_1 because $n m_n^{-1} \gg v_n = n m_n^{-2} B_1/3$. This, in conjunction with the fact that $B_1 m_n^{-1} \leq \lambda_{\min}\left(\Phi_l\right) \leq B_2 m_n^{-1}$ by Lemma A.3, implies

$$\Pr\left(\left|\lambda_{\min}\left(\Phi_{n,l}\right) - \lambda_{\min}\left(\Phi_{X_l}\right)\right| \geq \lambda_{\min}\left(\Phi_{X_l}\right)/3\right) \leq 2m_n^2 \exp\left\{-C_1 n m_n^{-3}\right\}. \qquad \text{(S1.6)}$$

Note that for two positive random variables a and b, $|a-b| \geq b/2$ is equivalent to $\{a - 3b/2 \geq 0$ or $a - b/2 \leq 0\}$, which is equivalent to

$$b^{-1} - a^{-1} \geq b^{-1}/3 \quad \text{or} \quad b^{-1} - a^{-1} \leq -b^{-1}.$$

Therefore, $\{|a-b| \geq b/2\}$ implies $\left\{\left|b^{-1} - a^{-1}\right| \geq b^{-1}/3\right\}$, and

$$\Pr\left(\{|a-b| \geq b/2\}\right) \leq \Pr\left(\left|b^{-1} - a^{-1}\right| \geq b^{-1}/3\right).$$

Taking $a = \left\{\lambda_{\min}\left(\Phi_{n,l}\right)\right\}^{-1}$ and $b = \left\{\lambda_{\min}\left(\Phi_{X_l}\right)\right\}^{-1}$, we have

$$\begin{aligned}
&\Pr\left(\left|\left[\lambda_{\min}\left(\Phi_{n,l}\right)\right]^{-1} - \left[\lambda_{\min}\left(\Phi_{X_l}\right)\right]^{-1}\right| \geq \left\{\lambda_{\min}\left(\Phi_l\right)\right\}^{-1}/2\right) \\
&\leq \Pr\left\{\left|\lambda_{\min}\left(\Phi_{n,l}\right) - \lambda_{\min}\left(\Phi_{X_l}\right)\right| \geq \lambda_{\min}\left(\Phi_l\right)/3\right\} \\
&\leq 2m_n^2 \exp\left\{-C_2 n m_n^{-3}\right\},
\end{aligned}$$

where the last inequality holds by (S1.6). This shows the first part of the lemma.

If Assumption 7 also holds, then

$$\begin{aligned}
m_n^2 \exp\left\{-C_2 n m_n^{-3}\right\} &= \exp\left\{-C_2 n m_n^{-3} + 2\log m_n\right\} \\
&= \exp\left\{-C_2 n^{1-3B_m} + 2B_m \log n\right\} \leq \exp\left\{-C_3 n^{C_4}\right\}
\end{aligned}$$

for some $C_4 \in (0, 1-3B_m)$ and $C_3 \in (0, C_2]$. Then the second part of the lemma follows.

To complete the proof the lemma, we now show (S1.2) and (S1.1). To see equation (S1.2), note that for any vector $\boldsymbol{x} = (x_1, ..., x_{m_n})'$ with $\|\boldsymbol{x}\| = 1$,

$$\min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \min_{\|\boldsymbol{x}\|=1}\left(\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} + \boldsymbol{x}'\left(\boldsymbol{A}-\boldsymbol{B}\right)\boldsymbol{x}\right) \geq \min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} + \min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}'\left(\boldsymbol{A}-\boldsymbol{B}\right)\boldsymbol{x},$$

which is equivalent to

$$\lambda_{\min}\left(\boldsymbol{A}\right) \geq \lambda_{\min}\left(\boldsymbol{B}\right) + \lambda_{\min}\left(\boldsymbol{A}-\boldsymbol{B}\right) \text{ or equivalently } \lambda_{\min}\left(\boldsymbol{A}-\boldsymbol{B}\right) \leq \lambda_{\min}\left(\boldsymbol{A}\right) - \lambda_{\min}\left(\boldsymbol{B}\right).$$

Switching $\boldsymbol{A}$ and $\boldsymbol{B}$ yields $\lambda_{\min}(\boldsymbol{B}-\boldsymbol{A}) \leq \lambda_{\min}(\boldsymbol{B}) - \lambda_{\min}(\boldsymbol{A})$. Therefore

$$\lambda_{\min}(\boldsymbol{A}-\boldsymbol{B}) \leq \lambda_{\min}(\boldsymbol{A}) - \lambda_{\min}(\boldsymbol{B}) \leq -\lambda_{\min}(\boldsymbol{B}-\boldsymbol{A}),$$

and equation (S1.2) follows. For (S1.1), we have by Jensen inequality

$$|\lambda_{\max}(\boldsymbol{A})| = \left|\max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\right| \leq \|\boldsymbol{A}\|_\infty \max_{\|\boldsymbol{x}\|=1} \left(\sum_{j=1}^{m_n} |x_j|\right)^2 \leq m_n \|\boldsymbol{A}\|_\infty \max_{\|\boldsymbol{x}\|=1} \sum_{j=1}^{m_n} x_j^2 = m_n \|\boldsymbol{A}\|_\infty,$$

and similarly

$$|\lambda_{\min}(\boldsymbol{A})| = \left|\min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\right| \leq \|\boldsymbol{A}\|_\infty \min_{\|\boldsymbol{x}\|=1} \left(\sum_{j=1}^{m_n} |x_j|\right)^2 \leq m_n \|\boldsymbol{A}\|_\infty \min_{\|\boldsymbol{x}\|=1} \sum_{j=1}^{m_n} x_j^2 = m_n \|\boldsymbol{A}\|_\infty.$$

This completes the proof of the lemma. ■

Proof of Lemma A.5. Noting that $\hat{\sigma}_l^2 = n^{-1}\boldsymbol{u}_l'(I_n - \mathbb{X}_l(\mathbb{X}_l'\mathbb{X}_l)^{-1}\mathbb{X}_l')\boldsymbol{u}_l$, we have

$$\hat{\sigma}_l^2 - \sigma_l^2 = n^{-1}\boldsymbol{u}_l'\boldsymbol{u}_l - \sigma_l^2 - n^{-1}\boldsymbol{u}_l'\mathbb{X}_l\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l.$$

It follows that

$$\begin{aligned}\Pr\left(\left|\hat{\sigma}_l^2 - \sigma_l^2\right| \geq v_n\right) \leq{}& \Pr\left(\left|n^{-1}\boldsymbol{u}_l'\boldsymbol{u}_l - \sigma_l^2\right| \geq (1-\pi_1)\, v_n/n\right) \\ &+ \Pr\left(\left|n^{-1}\boldsymbol{u}_l'\mathbb{X}_l\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right| \geq \pi_1 v_n/n\right)\end{aligned} \tag{S1.7}$$

for any $\pi_1 \in (0,1)$.

We first bound the first term on the right hand side of (S1.7). By Assumption 6 and

$$\begin{aligned}U_l &= Y - P^{m_n}(X_l)'\boldsymbol{\beta}_l \\ &= \sum_{j=1}^{p^*} f_j^*(X_j) - P^{m_n}(X_l)'\boldsymbol{\beta}_l + \varepsilon\end{aligned}$$

with

$$\boldsymbol{\beta}_l = \left[E\left(P^{m_n}(X_l)\,P^{m_n}(X_l)'\right)\right]^{-1} E\left(P^{m_n}(X_l)\,Y\right),$$

we can say that U_l is just a uniformly bounded random term plus ε because elements in $\boldsymbol{\beta}_l$ are uniformly bounded due to Lemma A.3 and the uniform boundedness of $\sum_{j=1}^{p^*} f_j^*(X_j)$, and for each x only a finite elements of $P^{m_n}(x)$ are nonzero due to the usage of finite order B-splines. This implies that U_l shares the same tail behavior as ε and satisfies Assumption 4. Then, the conditions in Lemma A.2 hold for $n^{-1}\boldsymbol{u}_l'\boldsymbol{u}_l - \sigma_l^2$ with $\alpha = s/2$ by Assumption 4. Applying Lemma A.2 yields that for any $v_n \propto n^\lambda$ with $1/2 < \lambda \leq (1+s/2)/(2+s/2)$, $\pi < \pi_1$

$$\begin{aligned}\Pr\left(\left|n^{-1}\boldsymbol{u}_l'\boldsymbol{u}_l - \sigma_l^2\right| \geq (1-\pi_1)\, v_n/n\right) &\leq \exp\left[-(1-\pi)^2 v_n^2 \Big/ \left(2n\omega_l^4\right)\right] \\ &= \frac{C_1}{2}\exp\left(-C_2 v_n^2/n\right) = \frac{C_1}{2}\exp\left(-C_2 n^{C_3}\right)\end{aligned} \tag{S1.8}$$

with $C_1 = 2$, $C_2 = (1-\pi)^2 \big/ \left(2\omega_l^4\right)$, $C_3 = 2\lambda - 1$, and last line holds by the fact that $\lambda > 1/2$. For $v_n \propto n^\lambda$ with $\lambda > (1+s/2)/(2+s/2)$, Lemma A.2 implies

$$\Pr\left(\left|n^{-1}\boldsymbol{u}_l'\boldsymbol{u}_l - \sigma_l^2\right| \geq (1-\pi_1)\, v_n/n\right) \leq \frac{C_1}{2}\exp\left(-C_2 n^{C_3}\right), \tag{S1.9}$$

for some $C_1, C_2, C_3 > 0$.

We turn to the second term on the right hand side of (S1.7). Note that Lemma A.6 (to be shown below) still holds if we remove $\hat{\sigma}_l^2$ and σ_l^2. Then by Lemma A.6,

$$\Pr\left(\left|n^{-1}\boldsymbol{u}_l'\mathbb{X}_l\left(\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right| \geq \pi_1 v_n/n\right) \leq \frac{C_1}{2}\exp\left(-C_2 n^{C_3}\right) \tag{S1.10}$$

for some positive constants C_1, C_2 and C_3 for $v_n \propto n^\lambda$ with $\lambda > 1/2$.[5]

Combining (S1.7), (S1.8), (S1.9), and (S1.10) completes the proof. ■

Proof of Lemma A.6. We show the first part first. Note that

$$\left|\boldsymbol{u}_l'\mathbb{X}_l\left(\hat{\sigma}_l^2\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\mathbb{X}_l'\boldsymbol{u}_l\right| \leq \hat{\sigma}_l^{-2}\left\|\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\right\|\left\|n^{-1/2}\mathbb{X}_l'\boldsymbol{u}_l\right\|^2, \tag{S1.11}$$

and we bound the three terms on the right hand side of (S1.11) in turn. For the first term, with $v_n = \frac{1}{4}n\sigma_l^2$, Lemma A.5 implies

$$\Pr\left(\hat{\sigma}_l^{-2} > \frac{4}{3}\sigma_l^{-2}\right) \leq \frac{C_2}{2}\exp\left(-C_3 n^{C_4}\right), \tag{S1.12}$$

for some positive constants C_2, C_3, and C_4. For the second term, equations (A.12) and (A.13) and Assumption 7 imply that

$$\Pr\left(\left\|\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\right\| \geq \frac{4}{3}B_1^{-1}m_n\right) \leq 2m_n^2\exp\left(-C_3 n m_n^{-3}\right) \leq \frac{C_2}{2}\exp\left(-C_3 n^{C_4}\right), \tag{S1.13}$$

for some positive constants C_2, C_3, and C_4, and $C_4 \leq 1 - 3B_m$. For the third term,

$$\left\|n^{-1/2}\mathbb{X}_l'\boldsymbol{u}_l\right\|^2 = \sum_{j=1}^{m_n}\left\{\sum_{i=1}^{n} n^{-1/2}\phi_j\left(x_{li}\right)u_{li}\right\}^2.$$

As we discuss in the proof of Lemma A.5, U_l also satisfies Assumption 4. Because $\phi_j(x)$ is uniformly bounded for all j, $\phi_j(X_l)\,U_l$ also satisfies Assumption 4. So

$$\begin{aligned}
\Pr\left(\left\|n^{-1/2}\mathbb{X}_l'\boldsymbol{u}_l\right\|^2 \geq \left(\frac{4}{3}B_1^{-1}m_n\right)^{-1} v_n\right) &= \Pr\left(\sum_{j=1}^{m_n}\left\{\sum_{i=1}^{n} n^{-1/2}\phi_j\left(x_{li}\right)u_{li}\right\}^2 \geq \frac{3}{4}B_1 m_n^{-1}v_n\right) \\
&\leq \sum_{j=1}^{m_n}\Pr\left(\left\{n^{-1/2}\sum_{i=1}^{n}\phi_j\left(x_{li}\right)u_{li}\right\}^2 \geq \frac{1}{m_n}\frac{3}{4}B_1 m_n^{-1}v_n\right) \\
&= \sum_{j=1}^{m_n}\Pr\left(\left|\sum_{i=1}^{n}\phi_j\left(x_{li}\right)u_{li}\right| \geq \left(\frac{3}{4}B_1\right)^{1/2} m_n^{-1}v_n^{1/2}n^{1/2}\right)
\end{aligned}$$

[5] It holds no matter which part of Lemma A.6 we apply.

$$\leq m_n \exp\left(-(1-\pi)^2 \frac{3}{4} B_1 m_n^{-2} v_n n \Big/ \left[2n\left(C_8 m_n^{-1}\right)\right]\right)$$

$$\leq \exp\left(-(1-\pi)^2 \frac{3}{8} B_1 C_8^{-1} m_n^{-1} v_n + \log m_n\right) \tag{S1.14}$$

for any $\pi \in (0,1)$, where the second inequality holds by the first part of Lemma A.2 and Assumption 4 with $\alpha = s$ and the fact that

$$\max_j \left\{E\left[\phi_j(X_l)^2 U_l^2\right]\right\} \leq \max_j \left\{E\left[\phi_j(X_l)^2 E(U_l^2|X_l)\right]\right\} \leq C \max_j \left\{E\left[\phi_j(X_l)^2\right]\right\} \leq C_8 m_n^{-1}$$

for some constants C and C_8 by Assumption 6. Here the second inequality in the last displayed line follows from the fact that that U_l is ε plus a uniformly bounded term, and the last inequality holds by Lemma A.3.

By (S1.11), (S1.12), (S1.13), and (S1.14), we have

$$\begin{aligned}
\Pr\left(\left|\boldsymbol{u}_l' \mathbb{X}_l \left(\hat{\sigma}_l^2 \mathbb{X}_l' \mathbb{X}_l\right)^{-1} \mathbb{X}_l' \boldsymbol{u}_l\right| \geq \frac{4}{3}\sigma_l^{-2} v_n\right) &\leq \Pr\left(\hat{\sigma}_l^{-2}\left\|n^{-1/2}\mathbb{X}_l'\boldsymbol{u}_l\right\|^2 \left\|\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\right\| \geq \frac{4}{3}\sigma_l^{-2} v_n\right) \\
&\leq \Pr\left(\hat{\sigma}_l^{-2} > \frac{4}{3}\sigma_l^{-2}\right) \\
&+ \Pr\left(\left\|\left(n^{-1}\mathbb{X}_l'\mathbb{X}_l\right)^{-1}\right\| \geq \frac{4}{3} B_1^{-1} m_n\right) \\
&+ \Pr\left(\left\|n^{-1/2}\mathbb{X}_l'\boldsymbol{u}_l\right\|^2 \geq \left(\frac{4}{3} B_1^{-1} m_n\right)^{-1} v_n\right) \\
&\leq \exp\left(-C_1 m_n^{-1} v_n + \log m_n\right) + C_2 \exp\left(-C_3 n^{C_4}\right)
\end{aligned}$$

for positive constants $C_1 = (1-\pi)^2 \frac{3}{8} B_1 C_8^{-1}$, C_2, C_3 and C_4. This completes the proof of the first part of the lemma.

The second part holds similarly. The only difference is that we apply the second part of Lemma A.2 to the fourth line of equation (S1.14). ■

Proof of Lemma A.7. With Assumption 5', Stone (1985) implies that there exists $\breve{\boldsymbol{\beta}}_j$ such that

$$\sup_{x\in[0,1]} \left|P^{m_n}(x)' \breve{\boldsymbol{\beta}}_j - f_j^*(x)\right| = O\left(m_n^{-d}\right) \tag{S1.15}$$

for $j = 1, 2, ..., p^*$. We rewrite Y as

$$\begin{aligned}
Y &= \sum_{j=1}^{p^*} P^{m_n}(X_j)' \breve{\boldsymbol{\beta}}_j + \sum_{j=1}^{p^*} \left[f_j^*(X_j) - P^{m_n}(X_j)' \breve{\boldsymbol{\beta}}_j\right] + \varepsilon \\
&\equiv \sum_{j=1}^{p^*} P^{m_n}(X_j)' \breve{\boldsymbol{\beta}}_j + \breve{R}_n + \varepsilon. \qquad \text{(S1.16)}
\end{aligned}$$

Note that $\breve{\boldsymbol{\beta}}_j$ is similar to $\tilde{\boldsymbol{\beta}}_j$ defined Section A.3. We employ $\breve{\boldsymbol{\beta}}_j$ here (and only here) because $\breve{R}_n = O\left(m_n^{-d}\right)$ by equation (S1.15) instead of $R_n = O_P\left(m_n^{-d}\right)$ implied by equation (A.4). The stronger result on $\breve{R}_n$ facilitates our proof below; see, e.g., the inequality in equation (S1.21). Moreover, if

$\left\{E\left[f_{nj}^{*}\left(X_j\right)^2\right]\right\}^{1/2} \gtrsim \kappa_n \log\left(m_n\right)\left(m_n/n\right)^{1/2}$ for some j, $\left\{E\left[P^{m_n}\left(X_j\right)'\breve{\boldsymbol{\beta}}_j\right]^2\right\}^{1/2} \approx \left\{E\left[f_{nj}^{*}\left(X_j\right)^2\right]\right\}^{1/2}$ because $R_n = O_P\left(m_n^{-d}\right)$ and $\breve{R}_n = O\left(m_n^{-d}\right)$ and the bias term is asymptotically negligible. Thus, by the third part of Lemma A.3,

$$\left\|\breve{\boldsymbol{\beta}}_j\right\| \propto m_n^{1/2}\left\{E\left[f_{nj}^{*}\left(X_j\right)^2\right]\right\}^{1/2} \gtrsim \kappa_n m_n \log\left(m_n\right) n^{-1/2}. \tag{S1.17}$$

Let $\Phi_{\boldsymbol{Z}} = E[P^{m_n}(\boldsymbol{Z})P^{m_n}(\boldsymbol{Z})']$, $\Phi_{\boldsymbol{X}_1^b} = E[P^{m_n}(\boldsymbol{X}_1^b)P^{m_n}(\boldsymbol{X}_1^b)']$, and $\Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}} = E[P^{m_n}(\boldsymbol{X}_1^b) \cdot P^{m_n}(\boldsymbol{Z})']$. Substituting equation (S1.16) into $\boldsymbol{\eta}_1^b$, we have

$$\begin{aligned}
\boldsymbol{\eta}_1^b &= E\left[P^{m_n}\left(\boldsymbol{X}_1^b\right)\left(P^{m_n}\left(\boldsymbol{X}_1^b\right)'\breve{\boldsymbol{\beta}}_1^b - P^{m_n}(\boldsymbol{Z})'\Phi_{\boldsymbol{Z}}^{-1}E\left[P^{m_n}(\boldsymbol{Z})P^{m_n}\left(\boldsymbol{X}_1^b\right)'\breve{\boldsymbol{\beta}}_1^b\right]\right)\right] \\
&\quad + E\left[P^{m_n}\left(\boldsymbol{X}_1^b\right)\left(\breve{R}_n - P^{m_n}(\boldsymbol{Z})'\Phi_{\boldsymbol{Z}}^{-1}E\left[P^{m_n}(\boldsymbol{Z})\breve{R}_n\right]\right)\right] \\
&= \left(\Phi_{\boldsymbol{X}_1^b} - \Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}\Phi_{\boldsymbol{Z}}^{-1}\Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}'\right)\breve{\boldsymbol{\beta}}_1^b + \left\{E\left[P^{m_n}\left(\boldsymbol{X}_1^b\right)\breve{R}_n\right] - \Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}\Phi_{\boldsymbol{Z}}^{-1}E\left[P^{m_n}(\boldsymbol{Z})\breve{R}_n\right]\right\},
\end{aligned} \tag{S1.18}$$

where $\breve{\boldsymbol{\beta}}_1^b \equiv \left(\breve{\boldsymbol{\beta}}_1', \ldots, \breve{\boldsymbol{\beta}}_b'\right)'$, the terms associated with $P^{m_n}\left(\boldsymbol{X}_{b+1}^{p^*}\right)'\breve{\boldsymbol{\beta}}_{b+1}^{p^*}$ are dropped out because $\boldsymbol{X}_{b+1}^{p^*}$ are included in $\boldsymbol{Z}$, and the terms associated with ε have zero expectation due to Assumption 1. We analyze the above two terms in $\boldsymbol{\eta}_1^b$ one by one.

We first study the first term in equation (S1.18). Let $\Phi = \begin{bmatrix} \Phi_{\boldsymbol{X}_1^b} & \Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}} \\ \Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}' & \Phi_{\boldsymbol{Z}} \end{bmatrix}$. Noting that $\boldsymbol{X}_1^b$ and $\boldsymbol{Z}$ are signals or pseudo-signals, we have by Assumption 10 that

$$\lambda_{\min}(\Phi) \propto m_n^{-1} \text{ and } \lambda_{\max}(\Phi) \propto m_n^{-1}. \tag{S1.19}$$

This implies that $\lambda_{\max}\left(\Phi^{-1}\right) = \left[\lambda_{\min}(\Phi)\right]^{-1} \propto m_n$ and $\lambda_{\min}\left(\Phi^{-1}\right) = \left[\lambda_{\max}(\Phi)\right]^{-1} \propto m_n$, which along with the *inclusion principle* (e.g., Theorem 8.4.5 in Bernstein (2005)) further implies that

$$\lambda_{\min}\left(\Phi_{\boldsymbol{X}_1^b} - \Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}\Phi_{\boldsymbol{Z}}^{-1}\Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}'\right) = \left[\lambda_{\max}\left(\left(\Phi_{\boldsymbol{X}_1^b} - \Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}\Phi_{\boldsymbol{Z}}^{-1}\Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}'\right)^{-1}\right)\right]^{-1} \propto m_n^{-1},$$

because $(\Phi_{\boldsymbol{X}_1^b} - \Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}\Phi_{\boldsymbol{Z}}^{-1}\Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}')^{-1}$ is the leading principal submatrix of Φ^{-1}. Then

$$\left\|\left(\Phi_{\boldsymbol{X}_1^b} - \Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}\Phi_{\boldsymbol{Z}}^{-1}\Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}'\right)\breve{\boldsymbol{\beta}}_1^b\right\| \gtrsim m_n^{-1}\left\|\breve{\boldsymbol{\beta}}_1^b\right\| \gtrsim m_n^{-1}\left\|\breve{\boldsymbol{\beta}}_j\right\| \tag{S1.20}$$

for any $1 \leq j \leq b$.

For the second term in equation (S1.18), we write $E[P^{m_n}(\boldsymbol{X}_1^b)\breve{R}_n] - \Phi_{\boldsymbol{X}_1^b\boldsymbol{Z}}\Phi_{\boldsymbol{Z}}^{-1}E\left[P^{m_n}(\boldsymbol{Z})\breve{R}_n\right] \equiv T_{n1} - T_{n2}$. For T_{n1},

$$\|T_{n1}\| \leq \left\|E\left[\left|P^{m_n}\left(\boldsymbol{X}_1^b\right)\right|\left|\breve{R}_n\right|\right]\right\| \lesssim m_n^{-d}\left\|E\left[\left|P^{m_n}\left(\boldsymbol{X}_1^b\right)\right|\right]\right\| = m_n^{-d}\left(\sum_{l=1}^{b}\sum_{k=1}^{m_n}\left\{E\left|\phi_k\left(X_l\right)\right|\right\}^2\right)^{1/2}, \tag{S1.21}$$

where $\lesssim$ holds by equation (S1.15), and the equality holds by the definitions of $P^{m_n}(\boldsymbol{X}_1^b)$ and $\|\cdot\|$. Here, $|A| \equiv (|a_1|, \ldots, |a_l|)'$ for a vector $A = (a_1, \ldots, a_l)'$. By the property of B-splines in Lemma A.3), $E\left|\phi_k\left(X_l\right)\right| \propto m_n^{-1}$. Then

$$\|T_{n1}\| \lesssim m_n^{-d}\left(\sum_{l=1}^{b}\sum_{k=1}^{m_n}\left\{E\left|\phi_k\left(X_l\right)\right|\right\}^2\right)^{1/2} \propto m_n^{-d-1/2}, \tag{S1.22}$$

where we use the fact that b is finite. For T_{n2}, we first claim

$$\left\|\Phi_{\boldsymbol{X}_1^b \boldsymbol{Z}}\right\| \lesssim m_n^{-1}. \tag{S1.23}$$

Then by the submultiplicative property of $\|\cdot\|$,

$$\|T_{n2}\| \leq \left\|E\left[P^{m_n}(\boldsymbol{Z})\breve{R}_n\right]\right\| \left\|\Phi_{\boldsymbol{Z}}^{-1}\right\| \left\|\Phi_{\boldsymbol{X}_1^b \boldsymbol{Z}}\right\| \lesssim m_n^{-d-1/2} m_n m_n^{-1} = m_n^{-d-1/2}, \tag{S1.24}$$

where second inequality holds by (S1.22), (S1.23), and Assumption 10. Consequently, we have

$$\|T_{n1} - T_{n2}\| \lesssim m_n^{-d-1/2}. \tag{S1.25}$$

If we have $\left\{E\left[f_{nj}^*(X_j)^2\right]\right\}^{1/2} \gtrsim \kappa_n \log(m_n)(m_n/n)^{1/2}$ for some j, then by (S1.20) and (S1.17),

$$\left\|\left(\Phi_{\boldsymbol{X}_1^b} - \Phi_{\boldsymbol{X}_1^b \boldsymbol{Z}}\Phi_{\boldsymbol{Z}}^{-1}\Phi'_{\boldsymbol{X}_1^b \boldsymbol{Z}}\right)\breve{\boldsymbol{\beta}}_1^b\right\| \gtrsim m_n^{-1}\left\|\breve{\boldsymbol{\beta}}_j\right\| \propto m_n^{-1/2}\left\{E\left[f_{nj}^*(X_j)^2\right]\right\}^{1/2} \gtrsim \kappa_n \log(m_n) n^{-1/2}.$$

This, in conjunction with (S1.25) and the fact that $\kappa_n \log(m_n) n^{-1/2} \gg m_n^{-d-1/2}$ by Assumption 7, implies that $\|T_{n1} - T_{n2}\|$ is of smaller order than $\left\|(\Phi_{\boldsymbol{X}_1^b} - \Phi_{\boldsymbol{X}_1^b \boldsymbol{Z}}\Phi_{\boldsymbol{Z}}^{-1}\Phi'_{\boldsymbol{X}_1^b \boldsymbol{Z}})\breve{\boldsymbol{\beta}}_1^b\right\|$ and thus

$$\left\|\boldsymbol{\eta}_1^b\right\| \gtrsim \kappa_n \log(m_n)^{1/2} n^{-1/2}.$$

Now we show equation (S1.23). Recall that $\Phi = \begin{bmatrix} \Phi_{\boldsymbol{X}_1^b} & \Phi_{\boldsymbol{X}_1^b \boldsymbol{Z}} \\ \Phi'_{\boldsymbol{X}_1^b \boldsymbol{Z}} & \Phi_{\boldsymbol{Z}} \end{bmatrix}$. By equation (S1.19), we have

$$\|\Phi\| = \max_{\substack{\left\|(\boldsymbol{a}_1', \boldsymbol{a}_2')'\right\|=1,\\ \boldsymbol{a}_1 \in \mathbb{R}^{bm_n}, \boldsymbol{a}_2 \in \mathbb{R}^{\iota_n m_n}}} (\boldsymbol{a}_1', \boldsymbol{a}_2')\,\Phi \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{pmatrix} \propto m_n^{-1}. \tag{S1.26}$$

Noting that $(\boldsymbol{a}_1', \boldsymbol{a}_2')\,\Phi \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{pmatrix} = \boldsymbol{a}_1'\Phi_{\boldsymbol{X}_1^b}\boldsymbol{a}_1 + \boldsymbol{a}_2'\Phi_{\boldsymbol{Z}}\boldsymbol{a}_2 + 2\boldsymbol{a}_1'\Phi_{\boldsymbol{X}_1^b \boldsymbol{Z}}\boldsymbol{a}_2$, we have by the nonnegative definiteness of $\Phi_{\boldsymbol{X}_1^b}$ and $\Phi_{\boldsymbol{Z}}$,

$$\max_{\substack{\left\|(\boldsymbol{a}_1', \boldsymbol{a}_2')'\right\|=1,\\ \boldsymbol{a}_1 \in \mathbb{R}^{bm_n}, \boldsymbol{a}_2 \in \mathbb{R}^{\iota_n m_n}}} \left|\boldsymbol{a}_1'\Phi_{\boldsymbol{X}_1^b \boldsymbol{Z}}\boldsymbol{a}_2\right| \leq \max_{\substack{\left\|(\boldsymbol{a}_1', \boldsymbol{a}_2')'\right\|=1,\\ \boldsymbol{a}_1 \in \mathbb{R}^{bm_n}, \boldsymbol{a}_2 \in \mathbb{R}^{\iota_n m_n}}} (\boldsymbol{a}_1', \boldsymbol{a}_2')\,\Phi \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{pmatrix} \propto m_n^{-1}, \tag{S1.27}$$

Then

$$\begin{aligned}
\left\|\Phi_{\boldsymbol{X}_1^b \boldsymbol{Z}}\right\| &= \max_{\substack{\|\boldsymbol{a}_1\|=1, \left\|\boldsymbol{a}_2'\right\|=1,\\ \boldsymbol{a}_1 \in \mathbb{R}^{bm_n}, \boldsymbol{a}_2 \in \mathbb{R}^{\iota_n m_n}}} \left|\boldsymbol{a}_1'\Phi_{\boldsymbol{X}_1^b \boldsymbol{Z}}\boldsymbol{a}_2\right| \leq \max_{\substack{\left\|(\boldsymbol{a}_1', \boldsymbol{a}_2')'\right\| \leq \sqrt{2}\\ \boldsymbol{a}_1 \in \mathbb{R}^{bm_n}, \boldsymbol{a}_2 \in \mathbb{R}^{\iota_n m_n}}} \left|\boldsymbol{a}_1'\Phi_{\boldsymbol{X}_1^b \boldsymbol{Z}}\boldsymbol{a}_2\right| \\
&= 2 \max_{\substack{\left\|(\boldsymbol{a}_1', \boldsymbol{a}_2')'\right\|=1\\ \boldsymbol{a}_1 \in \mathbb{R}^{bm_n}, \boldsymbol{a}_2 \in \mathbb{R}^{\iota_n m_n}}} \left|\boldsymbol{a}_1'\Phi_{\boldsymbol{X}_1^b \boldsymbol{Z}}\boldsymbol{a}_2\right| \propto m_n^{-1}
\end{aligned}$$

where the first equality follows from Fact 9.11.2 in Bernstein (2005), and the first inequality holds because $\{\|\boldsymbol{a}_1\|=1, \|\boldsymbol{a}_2'\|=1\} \subset \{\|(\boldsymbol{a}_1', \boldsymbol{a}_2')\| \leq \sqrt{2}\}$. ■

Proof of Lemma A.8. As in the proof of the last lemma, let $\Phi_{\boldsymbol{Z}} = E\left[P^{m_n}(\boldsymbol{Z})\, P^{m_n}(\boldsymbol{Z})'\right]$. We prove the lemma by showing that

$$\Pr\left(\left|\left\|\left(n^{-1}\mathbb{Z}'\mathbb{Z}\right)^{-1}\right\| - \left\|\Phi_{\boldsymbol{Z}}^{-1}\right\|\right| > \left\|\Phi_{\boldsymbol{Z}}^{-1}\right\|/\,2\right) \leq 2m_n^2\iota_n^2 \exp\left\{-C_1 n m_n^{-3}\iota_n^{-1}\right\},$$

for a $(\iota_n+1)\times 1$ random vector $\boldsymbol{Z}$, where we have absorbed X_l into $\boldsymbol{Z}$. Assumption 10 applies to $\boldsymbol{Z}$. Because of this, the proof is essentially the same as that for Lemma A.4. The only difference is that the dimension of the matrix here is $m_n(\iota_n+1)\times m_n(\iota_n+1)$.

Following the proof in Lemma A.4 up to equation (S1.4), we have

$$\begin{aligned}
&\Pr\left(\left|\lambda_{\min}\left(n^{-1}\mathbb{Z}'\mathbb{Z}\right) - \lambda_{\min}\left(\Phi_{\boldsymbol{Z}}\right)\right| \geq m_n(\iota_n+1)\,v_n/n\right)\\
&\leq \Pr\left(\max\left\{\left|\lambda_{\min}\left(n^{-1}\mathbb{Z}'\mathbb{Z}-\Phi_{\boldsymbol{Z}}\right)\right|, \left|\lambda_{\min}\left(\Phi_{\boldsymbol{Z}}-n^{-1}\mathbb{Z}'\mathbb{Z}\right)\right|\right\} \geq m_n(\iota_n+1)\,v_n/n\right)\\
&\leq \Pr\left(\left\|n^{-1}\mathbb{Z}'\mathbb{Z}-\Phi_{\boldsymbol{Z}}\right\|_\infty \geq v_n/n\right) \leq 2m_n^2(\iota_n+1)^2\exp\left\{-v_n^2/2\left(B_4 n m_n^{-1}+v_n/3\right)\right\}.
\end{aligned}$$

Set $v_n = n m_n^{-2}(\iota_n+1)^{-1}B_{X1}/3$, the above inequality becomes

$$\Pr\left(\left|\lambda_{\min}\left(n^{-1}\mathbb{Z}'\mathbb{Z}\right) - \lambda_{\min}\left(\Phi_{\boldsymbol{Z}}\right)\right| \geq B_{X1}m_n^{-1}/3\right) \leq 2m_n^2\iota_n^2\exp\left\{-C_1 n m_n^{-3}\iota_n^{-1}\right\}$$

for some positive constant C_1.

If in addition Assumption 2' and 7 holds, $n m_n^{-3}\iota_n^{-1} \geq n m_n^{-3} p_n^{**-1} \propto n^{1-3B_m-B_{p^{**}}} = n^{C_4}$ and $C_4 > 0$. The conclusion is immediate. ■

Proof of Lemma A.9. First

$$\begin{aligned}
&\Pr\left(\lambda_{\max}\left\{\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2 n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\} \geq 2\sigma^{-2}B_{X1}^{-1}m_n\right) \qquad \text{(S1.28)}\\
&\leq \Pr\left(\hat{\sigma}_{l,\boldsymbol{Z}}^{-2} \geq \frac{4}{3}\sigma^{-2}\right) + \Pr\left(\lambda_{\max}\left\{\left(n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\} \geq \frac{3}{2}B_{X1}^{-1}m_n\right).
\end{aligned}$$

Taking $v_n = \frac{1}{4}n\sigma^2$ in Lemma A.11 yields

$$\Pr\left(\hat{\sigma}_{l,\boldsymbol{Z}}^{-2} \geq \frac{4}{3}\sigma^{-2}\right) = \Pr\left(\hat{\sigma}_{l,\boldsymbol{Z}}^{2} \leq \frac{3}{4}\sigma^{2}\right) \leq \frac{1}{2}C_1\exp(-C_2 n^{C_3}) \qquad \text{(S1.29)}$$

for some positive C_1, C_2, and C_3. Noting that $\left(\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}$ is the lower block diagonal of $\left(\begin{pmatrix}\mathbb{Z}'\\ \mathbb{X}_l'\end{pmatrix}\begin{pmatrix}\mathbb{Z} & \mathbb{X}_l\end{pmatrix}\right)^{-1}$, by Lemma A.8 and Assumption 10, we have

$$\Pr\left(\lambda_{\max}\left\{\left(n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\} \geq \frac{3}{2}B_{X1}^{-1}m_n\right) \leq 2m_n^2\iota_n^2\exp\left\{-C_1 n m_n^{-3}\iota_n^{-1}\right\}.$$

Since $\iota_n \leq p^{**}$, $n m_n^{-3}\iota_n \leq n^{1-3B_m+B_{p^{**}}} = n^{C_3}$ and $C_3 > 0$ by Assumption 2'. Then

$$\Pr\left(\lambda_{\max}\left\{\left(n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\} \geq \frac{3}{2}B_{X1}^{-1}m_n\right) \leq \frac{1}{2}C_1\exp(-C_2 n^{C_3}). \qquad \text{(S1.30)}$$

Combining (S1.28), (S1.29), and (S1.30) yields

$$\Pr\left(\lambda_{\max}\left\{\left(\hat{\sigma}^2_{l,\boldsymbol{Z}} n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\} \geq 2\sigma^{-2}B_{X1}^{-1}m_n\right) \leq C_1\exp(-C_2 n^{C_3}).$$

For the second part, we first make the following decomposition:

$$\begin{aligned}
&\Pr\left(\lambda_{\min}\left\{\left(\hat{\sigma}^2_{l,\boldsymbol{Z}} n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\} \leq \frac{1}{4}\sigma^{-2}B_{X2}^{-1}m_n\right) \\
&\leq \Pr\left(\hat{\sigma}^{-2}_{l,\boldsymbol{Z}} \leq \frac{1}{2}\sigma^{-2}\right) + \Pr\left(\lambda_{\min}\left\{\left(n^{-1}\mathbb{X}_l' M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\} \leq \frac{1}{2}\sigma^{-2}B_{X2}^{-1}m_n\right).
\end{aligned}$$

Then we apply Lemma A.11 and Lemma A.8 to the two terms on the right hand side of the last displayed equation. The conclusion follows as in the proof of the first part. ■

Proof of Lemma A.10. We deal with $n^{-1/2}\left(\boldsymbol{u}'_{X_l,\boldsymbol{Z}}\boldsymbol{u}_{Y,\boldsymbol{Z}} - \boldsymbol{\eta}_{l,\boldsymbol{Z}}\right)$ first. Note

$$n^{-1/2}\left(\boldsymbol{u}'_{X_l,\boldsymbol{Z}}\boldsymbol{u}_{Y,\boldsymbol{Z}} - \boldsymbol{\eta}_{l,\boldsymbol{Z}}\right) = n^{-1/2}\sum_{i=1}^{n}\left(\boldsymbol{u}_{X_l,\boldsymbol{Z},i}u_{Y,\boldsymbol{Z},i} - \boldsymbol{\eta}_{l,\boldsymbol{Z}}\right)$$

Denote the j-th element of $\boldsymbol{U}_{X_l,\boldsymbol{Z}}$ by $U_{\phi_j(X_l),\boldsymbol{Z}}$. By equation (A.9) and Lemma A.3, $E\left(U^2_{\phi_j(X_l),\boldsymbol{Z}}\right) \propto m_n^{-1}$, and $U_{\phi_j(X_l),\boldsymbol{Z}}$ is uniformly bounded due to the uniform boundedness of $\phi_j(X_l)$.

By equation (A.9),

$$U_{Y,\boldsymbol{Z}} = \sum_{j=1}^{p^*} f_j^*(X_j) + \varepsilon - \boldsymbol{\gamma}'_{Y,\boldsymbol{Z}}P^{m_n}(\boldsymbol{Z}),$$

where $\boldsymbol{\gamma}_{Y,\boldsymbol{Z}} \equiv \Phi_{\boldsymbol{Z}}^{-1}E\left[P^{m_n}(\boldsymbol{Z})Y\right]$. The rank conditions in Assumption 10 and uniform boundedness of $\sum_{j=1}^{p^*} f_j^*(X_j)$ imply that each element in $\boldsymbol{\gamma}_{Y,\boldsymbol{Z}}$ is uniformly bounded. Note p^* is fixed and the dimension of $\boldsymbol{\gamma}_{Y,\boldsymbol{Z}}$ is $\iota_n m_n$, and thus

$$U_{Y,\boldsymbol{Z}} = C_{Y,\boldsymbol{Z}} + \varepsilon,$$

with $C_{Y,\boldsymbol{Z}} \equiv \sum_{j=1}^{p^*} f_j^*(X_j) - \boldsymbol{\gamma}'_{Y,\boldsymbol{Z}}P^{m_n}(\boldsymbol{Z})$ and $|C_{Y,\boldsymbol{Z}}| \leq C\iota_n m_n$ for a positive C. The j-th element of $\boldsymbol{u}'_{X_l,\boldsymbol{Z}}\boldsymbol{u}_{Y,\boldsymbol{Z}} - \boldsymbol{\eta}_{l,\boldsymbol{Z}}$ is

$$\sum_{i=1}^{n}\left(u_{\phi_j(X_l),\boldsymbol{Z},i}u_{Y,\boldsymbol{Z},i} - \boldsymbol{\eta}_{l,\boldsymbol{Z},\boldsymbol{j}}\right) = \sum_{i=1}^{n}\left(u_{\phi_j(X_l),\boldsymbol{Z},i}C_{Y,\boldsymbol{Z},i} - \boldsymbol{\eta}_{l,\boldsymbol{Z},\boldsymbol{j}}\right) + \sum_{i=1}^{n}u_{\phi_j(X_l),\boldsymbol{Z},i}\varepsilon_i.$$

Since $u_{\phi_j(X_l),\boldsymbol{Z},i}$ is uniformly bounded, $\left|u_{\phi_j(X_l),\boldsymbol{Z},i}C_{Y,\boldsymbol{Z},i}\right| \leq C\iota_n m_n$ for a positive C, and $u_{\phi_j(X_l),\boldsymbol{Z},i}\varepsilon_i$ satisfies the tail restriction in Assumption 4. Apply the inequalities in Lemmas A.1 on the first term in the above, we obtain

$$\begin{aligned}
&\Pr\left(\left|\sum_{i=1}^{n}\left(u_{\phi_j(X_l),\boldsymbol{Z},i}u_{Y,\boldsymbol{Z},i} - \boldsymbol{\eta}_{l,\boldsymbol{Z},\boldsymbol{j}}\right)\right| > v_n\right) \\
&\leq \Pr\left(\left|\sum_{i=1}^{n}\left(u_{\phi_j(X_l),\boldsymbol{Z},i}C_{Y,\boldsymbol{Z},i} - \boldsymbol{\eta}_{l,\boldsymbol{Z},\boldsymbol{j}}\right)\right| > \frac{v_n}{2}\right) + \Pr\left(\left|\sum_{i=1}^{n}u_{\phi_j(X_l),\boldsymbol{Z},i}\varepsilon_i\right| > \frac{v_n}{2}\right) \\
&\leq 2\exp\left\{-v_n^2/\left(C\left(nm_n^{-1} + \iota_n m_n v_n\right)\right)\right\} + \Pr\left(\left|\sum_{i=1}^{n}u_{\phi_j(X_l),\boldsymbol{Z},i}\varepsilon_i\right| > \frac{v_n}{2}\right). \qquad \text{(S1.31)}
\end{aligned}$$

Using the above identity, we have

$$\begin{aligned}
&\Pr\left(\left\|n^{-1/2}\sum_{i=1}^{n}\left(\boldsymbol{u}_{X_l,\boldsymbol{Z},i}u_{Y,\boldsymbol{Z},i}-\boldsymbol{\eta}_{l,\boldsymbol{Z}}\right)\right\|^2\geq C_1m_n^{-1}v_n\right)\\
&=\Pr\left(\sum_{j=1}^{m_n}\left\{n^{-1/2}\sum_{i=1}^{n}\left(u_{\phi_j(X_l),\boldsymbol{Z},i}u_{Y,\boldsymbol{Z},i}-\boldsymbol{\eta}_{l,\boldsymbol{Z},j}\right)\right\}^2\geq C_1m_n^{-1}v_n\right)\\
&\leq\sum_{j=1}^{m_n}\Pr\left(\left|\sum_{i=1}^{n}\left(u_{\phi_j(X_l),\boldsymbol{Z},i}u_{Y,\boldsymbol{Z},i}-\boldsymbol{\eta}_{l,\boldsymbol{Z},j}\right)\right|\geq C_1n^{1/2}m_n^{-1}v_n^{1/2}\right)\\
&\leq\sum_{j=1}^{m_n}\left\{\Pr\left(\left|\sum_{i=1}^{n}\left(u_{\phi_j(X_l),\boldsymbol{Z},i}C_{Y,\boldsymbol{Z},i}-\boldsymbol{\eta}_{l,\boldsymbol{Z},j}\right)\right|\geq \frac{1}{2}C_1n^{1/2}m_n^{-1}v_n^{1/2}\right)\right.\\
&\qquad\left.+\Pr\left(\left|\sum_{i=1}^{n}u_{\phi_j(X_l),\boldsymbol{Z},i}\varepsilon_i\right|\geq\frac{1}{2}C_1n^{1/2}m_n^{-1}v_n^{1/2}\right)\right\}\\
&\leq\begin{cases}2m_n\exp\left\{-nm_n^{-2}v_n/[C_2(nm_n^{-1}+\iota_nn^{1/2}v_n^{1/2})]\right\}+m_n\exp\left(-C_3m_n^{-1}v_n\right) & \text{if } v_n\lesssim n^{s/(s+2)}m_n^2\\ 2m_n\exp\left\{-nm_n^{-2}v_n/[C_2(nm_n^{-1}+\iota_nn^{1/2}v_n^{1/2})]\right\}+C_4\exp\left(-C_5n^{C_6}\right) & \text{if } v_n\gtrsim n^{s/(s+2)}m_n^2\end{cases},
\end{aligned}\tag{S1.32}$$

for any positive constant C_1 and some positive constants C_2, C_3, C_4 and C_5, where the last inequality follows from the same arguments as used to obtain (S1.14) along with applying equation (S1.31).

We turn to $n^{-1/2}\boldsymbol{u}'_{X_l,\boldsymbol{Z}}\mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}$. Note this is a $m_n\times 1$ vector whose j-th element is $n^{-1/2}\boldsymbol{u}'_{\phi_j(X_l),\boldsymbol{Z}}\mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}$ with the $n\times 1$ vector $\boldsymbol{u}_{\phi_j(X_l),\boldsymbol{Z}}$ being an collecting of the n observations for $U_{\phi_j(X_l),\boldsymbol{Z}}$. Then

$$\begin{aligned}
&\Pr\left(\left\|n^{-1/2}\boldsymbol{u}'_{X_l,\boldsymbol{Z}}\mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}\right\|^2\geq C_1m_n^{-1}v_n\right)\\
&=\Pr\left(\sum_{j=1}^{m_n}\left(n^{-1/2}\boldsymbol{u}'_{\phi_j(X_l),\boldsymbol{Z}}\mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}\right)^2\geq C_1m_n^{-1}v_n\right)\\
&\leq\sum_{j=1}^{m_n}\Pr\left(\left(n^{-1/2}\boldsymbol{u}'_{\phi_j(X_l),\boldsymbol{Z}}\mathbb{Z}\left(n^{-1}\mathbb{Z}'\mathbb{Z}\right)^{-1}n^{-1/2}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}\right)^2\geq C_1nm_n^{-2}v_n\right)\\
&\leq\sum_{j=1}^{m_n}\Pr\left(\left\|n^{-1/2}\mathbb{Z}'\boldsymbol{u}_{\phi_j(X_l),\boldsymbol{Z}}\right\|^2\left\|n^{-1/2}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}\right\|^2\lambda_{\max}\left\{\left(n^{-1}\mathbb{Z}'\mathbb{Z}\right)^{-1}\right\}^2\geq C_1nm_n^{-2}v_n\right).
\end{aligned}\tag{S1.33}$$

We deal with the three random terms in the last line in turn. First, by Assumptions 10 and 2' and Lemma A.8, we have

$$\Pr\left(\lambda_{\max}\left\{\left(n^{-1}\mathbb{Z}'\mathbb{Z}\right)^{-1}\right\}^2\geq\left(\frac{4}{3}B_{X1}^{-1}m_n\right)^2\right)\leq C_6\exp\left(-C_7n^{C_8}\right)\tag{S1.34}$$

for some positive constants C_6, C_7 and C_8. For the second term,

$$\Pr\left(\left\|n^{-1/2}\mathbb{Z}'\boldsymbol{u}_{\phi_j(X_l),\boldsymbol{Z}}\right\|^2\geq C_1^{1/2}\frac{3}{4}B_{X1}n^{1/2}m_n^{-2}v_n^{1/2}\right)$$

$$= \Pr\left(\sum_{k_1=1}^{\iota_n}\sum_{k_2=1}^{m_n}\left(\sum_{i=1}^{n} n^{-1/2}\phi_{k_2}\left(z_{k_1,i}\right)u_{\phi_j(X_l),\boldsymbol{Z},i}\right)^2 \geq C_1^{1/2}\frac{3}{4}B_{X1}n^{1/2}m_n^{-2}v_n^{1/2}\right)$$

$$\leq \sum_{k_1=1}^{\iota_n}\sum_{k_2=1}^{m_n}\Pr\left(\left(\sum_{i=1}^{n} n^{-1/2}\phi_{k_2}\left(z_{k_1,i}\right)u_{\phi_j(X_l),\boldsymbol{Z},i}\right)^2 \geq C_1^{1/2}\frac{3}{4}B_{X1}n^{1/2}m_n^{-3}\iota_n^{-1}v_n^{1/2}\right)$$

$$= \sum_{k_1=1}^{\iota_n}\sum_{k_2=1}^{m_n}\Pr\left(\left|\sum_{i=1}^{n}\phi_{k_2}\left(z_{k_1,i}\right)u_{\phi_j(X_l),\boldsymbol{Z},i}\right| \geq C_9 n^{3/4}m_n^{-3/2}\iota_n^{-1/2}v_n^{1/4}\right)$$

$$\leq \begin{cases} m_n\iota_n\exp\left(-C_{10}n^{1/2}m_n^{-3/2}\iota_n^{-1}\left(m_n^{-1}v_n\right)^{1/2}\right) & \text{if } v_n \lesssim n^{(s-2)/[4(s+2)]}m_n^{3/2}\iota_n^2 \\ C_{11}\exp\left(-C_{12}n^{C_{13}}\right) & \text{if } v_n \gtrsim n^{(s-2)/[4(s+2)]}m_n^{3/2}\iota_n^2 \end{cases}, \tag{S1.35}$$

for some positive constants C_9, C_{10}, C_{11} and C_{12}, where in the last line we apply Lemma A.2 and we use the fact that $\mathrm{var}\left(\phi_{k_2}\left(z_{k_1,i}\right)u_{\phi_j(X_l),\boldsymbol{Z},i}\right) \propto m_n^{-1}$. For the third term, similarly,

$$\Pr\left(\left\|n^{-1/2}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}\right\|^2 \geq C_1^{1/2}\frac{3}{4}B_{X1}n^{1/2}m_n^{-2}v_n^{1/2}\right)$$

$$= \Pr\left(\sum_{k_1=1}^{\iota_n}\sum_{k_2=1}^{m_n}\left(\sum_{i=1}^{n} n^{-1/2}\phi_{k_2}\left(z_{k_1,i}\right)u_{Y,\boldsymbol{Z},i}\right)^2 \geq C_1^{1/2}\frac{3}{4}B_{X1}n^{1/2}m_n^{-2}v_n^{1/2}\right)$$

$$\leq \begin{cases} m_n\iota_n\exp\left(-C_{10}n^{1/2}m_n^{-3/2}\iota_n^{-1}\left(m_n^{-1}v_n\right)^{1/2}\right) & \text{if } v_n \lesssim n^{(s-2)/[4(s+2)]}m_n^{3/2}\iota_n^2 \\ C_{11}\exp\left(-C_{12}n^{C_{13}}\right) & \text{if } v_n \gtrsim n^{(s-2)/[4(s+2)]}m_n^{3/2}\iota_n^2 \end{cases}, \tag{S1.36}$$

where without loss of generality we use the same constants in equation (S1.35).

Equation (S1.33) implies

$$\Pr\left(\left\|n^{-1/2}\boldsymbol{u}'_{X_l,\boldsymbol{Z}}\mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}\right\|^2 \geq C_1 m_n^{-1}v_n\right)$$

$$\leq \sum_{j=1}^{m_n}\left\{\Pr\left(\lambda_{\max}\left\{\left(n^{-1}\mathbb{Z}'\mathbb{Z}\right)^{-1}\right\}^2 \geq \left(\frac{4}{3}B_{X1}^{-1}m_n\right)^2\right)\right.$$

$$+\Pr\left(\left\|n^{-1/2}\mathbb{Z}'\boldsymbol{u}_{\phi_j(X_l),\boldsymbol{Z}}\right\|^2 \geq C_1^{1/2}\frac{3}{4}B_{X1}n^{1/2}m_n^{-2}v_n^{1/2}\right)$$

$$\left.+\Pr\left(\left\|n^{-1/2}\mathbb{Z}'\boldsymbol{u}_{\phi_j(X_l),\boldsymbol{Z}}\right\|^2 \geq C_1^{1/2}\frac{3}{4}B_{X1}n^{1/2}m_n^{-2}v_n^{1/2}\right)\right\}.$$

Substituting the results of equations (S1.34), (S1.35), and (S1.36) into the above yields

$$\Pr\left(\left\|n^{-1/2}\boldsymbol{u}'_{X_l,\boldsymbol{Z}}\mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}\right\|^2 \geq C_1 m_n^{-1}v_n\right) \tag{S1.37}$$

$$\leq \begin{cases} \sum_{j=1}^{m_n}\left[C_6\exp\left(-C_7n^{C_8}\right)+2m_n\iota_n\exp\left(-C_{10}n^{1/2}m_n^{-3/2}\iota_n^{-1}\left(m_n^{-1}v_n\right)^{1/2}\right)\right] & \text{if } v_n \lesssim n^{\frac{s-2}{4(s+2)}}m_n^{3/2}\iota_n^2 \\ \sum_{j=1}^{m_n}\left[C_6\exp\left(-C_7n^{C_8}\right)+2C_{11}\exp\left(-C_{12}n^{C_{13}}\right)\right] & \text{if } v_n \gtrsim n^{\frac{s-2}{4(s+2)}}m_n^{3/2}\iota_n^2 \end{cases}$$

$$\leq \begin{cases} C_{13}\exp\left(-C_{14}n^{C_{15}}\right)+m_n^2\iota_n\exp\left(-C_{16}n^{1/2}m_n^{-3/2}\iota_n^{-1}\left(m_n^{-1}v_n\right)^{1/2}\right) & \text{if } v_n \lesssim n^{\frac{s-2}{4(s+2)}}m_n^{3/2}\iota_n^2 \\ C_{17}\exp\left(-C_{18}n^{C_{19}}\right) & \text{if } v_n \gtrsim n^{\frac{s-2}{4(s+2)}}m_n^{3/2}\iota_n^2 \end{cases},$$

for some positive constants $C_{13},\dots,C_{19}$.

Note that when $\upsilon_n=\varsigma_n$, $n^{1/2}m_n^{-3/2}\iota_n^{-1}\left(m_n^{-1}\varsigma_n\right)^{1/2}\gg n^{1/2-1/6B_m-B_{p^{**}}}=n^{C_{22}}$ for a positive C_{22}. C_{22} exists due to Assumption 2'. Then,

$$\begin{aligned}&\Pr\left(\left\|n^{-1/2}\boldsymbol{u}_{X_l,\boldsymbol{Z}}'\mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}\right\|^2\geq C_1m_n^{-1}\varsigma_n\right)\\&\leq C_{20}\exp\left(-C_{21}n^{C_{22}}\right)\end{aligned}\tag{S1.38}$$

for some positive constants C_{20},C_{21}, and C_{22}, due to equation (S1.37) and the above observation. Therefore, by equations (S1.32) and (S1.38),

$$\begin{aligned}&\Pr\left(\left\|n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right\|^2\geq C_1m_n^{-1}\varsigma_n\right)\\&=\Pr\left(\left\|n^{-1/2}\left(\boldsymbol{u}_{X_l,\boldsymbol{Z}}'\boldsymbol{u}_{Y,\boldsymbol{Z}}-\boldsymbol{\eta}_{l,\boldsymbol{Z}}\right)-n^{-1/2}\boldsymbol{u}_{X_l,\boldsymbol{Z}}'\mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}\right\|^2\geq C_1m_n^{-1}\varsigma_n\right)\\&\leq\Pr\left(\left\|n^{-1/2}\left(\boldsymbol{u}_{X_l,\boldsymbol{Z}}'\boldsymbol{u}_{Y,\boldsymbol{Z}}-\boldsymbol{\eta}_{l,\boldsymbol{Z}}\right)\right\|^2\geq\frac{1}{2}C_1m_n^{-1}\varsigma_n\right)+\Pr\left(\left\|n^{-1/2}\boldsymbol{u}_{X_l,\boldsymbol{Z}}'\mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'\boldsymbol{u}_{Y,\boldsymbol{Z}}\right\|^2\geq\frac{1}{2}C_1m_n^{-1}\varsigma_n\right)\\&\leq 2\exp\left(-C_2m_n^{-1}\varsigma_n+\log m_n\right)+C_{20}\exp\left(-C_{21}n^{C_{22}}\right)\\&\leq n^{-M}+C_{20}\exp\left(-C_{21}n^{C_{22}}\right),\end{aligned}$$

for an arbitrarily large constant M, where in the fourth line we use the fact that the two terms in equation are of the same order due to $\iota_n n_n^{1/2}\varsigma_n^{1/2}\ll nm_n^{-1}$ by Assumption 2'. This completes the proof of the lemma. ■

Proof of Lemma A.11. To save notations, we absorb X_l into $\boldsymbol{Z}$ which now becomes a $(\iota_n+1)\times 1$ vector. Assumption 10 holds for $\boldsymbol{Z}$. Recall $\Phi_{\boldsymbol{Z}}=E\left[P^{m_n}(\boldsymbol{Z})P^{m_n}(\boldsymbol{Z})'\right]$ and $M_{\mathbb{Z}}=I_n-\mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'$. By definition,

$$U=Y-P^{m_n}(\boldsymbol{Z})'\Phi_{\boldsymbol{Z}}^{-1}E\left[P^{m_n}(\boldsymbol{Z})Y\right].$$

Let $u_i=Y_i-P^{m_n}(\boldsymbol{z}_i)'\Phi_{\boldsymbol{Z}}^{-1}E\left[P^{m_n}(\boldsymbol{Z})Y\right]$ and $\boldsymbol{u}=(u_1,\dots,u_n)$. By the definition of $\hat{\sigma}_{l,\boldsymbol{Z}}^2$,[6]

$$\hat{\sigma}_{l,\boldsymbol{Z}}^2=n^{-1}\boldsymbol{u}'M_{\mathbb{Z}}\boldsymbol{u}.$$

Then we are able to reach the same conclusion as in the proof of Lemma A.5, with the help of Lemma A.10. ■

Proof of Lemma A.12. Note that

$$\begin{aligned}\Pr\left(\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}'\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2\mathbb{X}_l'M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\geq\varsigma_n\right)&\leq\Pr\left(\lambda_{\max}\left\{\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2n^{-1}\mathbb{X}_l'M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\}\left\|n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right\|^2\geq\varsigma_n\right)\\&\leq\Pr\left(\lambda_{\max}\left\{\left(\hat{\sigma}_{l,\boldsymbol{Z}}^2n^{-1}\mathbb{X}_l'M_{\mathbb{Z}}\mathbb{X}_l\right)^{-1}\right\}\geq 2\sigma^{-2}B_{X1}^{-1}m_n\right)\\&\quad+\Pr\left(\left\|n^{-1/2}\tilde{\boldsymbol{u}}_{l,\boldsymbol{Z}}\right\|^2\geq\left(2\sigma^{-2}\right)^{-1}B_{X1}m_n^{-1}\varsigma_n\right).\end{aligned}$$

[6] We defined $\hat{\sigma}_{l,\boldsymbol{Z}}^2$ based on $\hat{u}_i$ from partitioned regressions. This $\hat{u}_i$ is the same as the residual from the joint regression.

Then we can reach the conclusion by applying Lemma A.9 and the last part of Lemma A.10. ■

Proof of Lemma A.13. We prove the lemma by mathematical induction. First, we show that the conclusion holds when $k = 1$, viz., for the first stage. We bound the probability of $\mathcal{N}_1$ (one or more noise variables are selected at stage 1) as

$$\Pr\left(\mathcal{N}_1\right) = \Pr\left(\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{L}_{l,1}\right) = \Pr\left(\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,1}\right) \leq \sum_{l=p^*+p^{**}+1}^{p_n} \Pr\left(\mathcal{B}_{l,1}\right). \tag{S1.39}$$

By the definition in Table 2, $\theta_l \lesssim \log\left(m_n\right)^{1/2}\left(m_n/n\right)^{1/2}$ for noise variables, viz., for $p^* + p^{**} + 1 \leq l \leq p_n$. Then we are able to apply the first part of Proposition 2.1 to bound $\Pr\left(\mathcal{B}_{l,1}\right)$ for each $l \in [p^* + p^{**} + 1, p_n]$ to obtain

$$\Pr\left(\mathcal{B}_{l,1}\right) \leq n^{-M} + C_1 \exp\left(-C_2 n^{C_3}\right)$$

for any arbitrarily large constant M and some positive constants C_1, C_2 and C_3. Then by Assumption 2 and (S1.39), we have

$$\begin{aligned}\Pr\left(\mathcal{N}_1\right) &\leq \sum_{l=p^*+p^{**}+1}^{p_n} \left(n^{-M} + C_1 \exp\left(-C_2 n^{C_3}\right)\right) \\ &\leq p_n n^{-M} + p_n C_1 \exp\left(-C_2 n^{C_3}\right) \\ &\leq n^{-M_1} + C_7 \exp\left(-C_5 n^{C_6}\right),\end{aligned} \tag{S1.40}$$

where the last equality holds by some $0 < M_1 \leq M - B_p$ and some constants C_5, C_6 and C_7. Since M is arbitrarily large and B_p is a fixed number, M_1 can also be arbitrarily large. Notice that $\Pr\left(\mathcal{D}_1\right) = 1 - \Pr\left(\mathcal{N}_1\right)$, we get

$$\Pr\left(\mathcal{D}_1\right) \geq 1 - n^{-M_1} - C_7 \exp\left(-C_5 n^{C_6}\right).$$

We now show the conclusion for $k = 2$. Notice that $\mathcal{N}_2 = \mathcal{N}_1 \cup \left\{\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,2}\right\}$. That is, some noise variables selected up to and including stage 2 are either selected at stage 1 or at stage 2. To apply Proposition 3.1, we need $\mathcal{D}_1$: the pre-selected variables are either signals or pseudo-signals at stage 2. In view of this and the fact that $\mathcal{D}_1 = \mathcal{N}_1^c$, we have

$$\begin{aligned}\Pr\left(\mathcal{N}_2\right) &= \Pr\left(\mathcal{N}_1 \cup (\left\{\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,2}\right\} \cap \mathcal{D}_1)\right) \\ &\leq \Pr\left(\mathcal{N}_1\right) + \Pr\left(\left\{\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,2}\right\} \cap \mathcal{D}_1\right) \\ &= \Pr\left(\mathcal{N}_1\right) + \Pr\left(\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,2} \cap \mathcal{D}_1\right) \\ &\leq \Pr\left(\mathcal{N}_1\right) + \Pr\left(\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,2} | \mathcal{D}_1\right) \Pr\left(\mathcal{D}_1\right) \\ &\leq \Pr\left(\mathcal{N}_1\right) + \Pr\left(\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,2} | \mathcal{D}_1\right).\end{aligned} \tag{S1.41}$$

where the first equality follows from the fact $\Pr\left(A \cup B\right) = \Pr\left(A \cup \left(B \cap A^c\right)\right)$ for any two events A and B. Since the second term in (S1.41) is conditional on $\mathcal{D}_1$, the pre-selected variables in $\boldsymbol{Z}_{(1)}$ are either signals or pseudo signals. By Assumption 11, the net effect of the noise variable X_l on Y with $\boldsymbol{Z}_{(1)}$ as pre-selected variables satisfies $\boldsymbol{\theta}_{l,\boldsymbol{Z}_{(1)}} \lesssim \log(m_n)^{1/2} m_n^{1/2} n^{-1/2}$. Then we are able to apply the first part of Proposition 3.1 to obtain

$$\begin{aligned}\Pr\left(\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,2} | \mathcal{D}_1\right) &\leq \sum_{l=p^*+p^{**}+1}^{p_n} \Pr\left(\mathcal{B}_{l,2} | \mathcal{D}_1\right) \leq \sum_{l=p^*+p^{**}+1}^{p_n} \left(n^{-M} + C_1 \exp\left(-C_2 n^{C_3}\right)\right) \\ &\leq n^{-M_1} + C_7 \exp\left(-C_5 n^{C_6}\right), \end{aligned} \tag{S1.42}$$

where without loss of generality we use the same constants as in equation (S1.40). Combining (S1.40)–(S1.42) yields

$$\Pr\left(\mathcal{N}_2\right) \leq 2n^{-M_1} + 2C_7 \exp\left(-C_5 n^{C_6}\right)$$

and

$$\Pr\left(\mathcal{D}_2\right) = 1 - \Pr\left(\mathcal{N}_2\right) \geq 1 - 2n^{-M_1} - 2C_7 \exp\left(-C_5 n^{C_6}\right).$$

Now suppose we have shown the results for stage $k-1$ with $k \geq 3$, and obtained

$$\Pr\left(\mathcal{N}_{k-1}\right) \leq (k-1) n^{-M_1} + (k-1) C_7 \exp\left(-C_5 n^{C_6}\right). \tag{S1.43}$$

Note that $\mathcal{N}_k = \mathcal{N}_{k-1} \cup \left\{\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,k}\right\}$, viz., some noise variables selected up to and including stage k are either selected at stage k or before stage k. Noting that $\mathcal{D}_{k-1} = \mathcal{N}_{k-1}^c$ and following the derivation of equation (S1.41), we have

$$\begin{aligned}\Pr\left(\mathcal{N}_k\right) &= \Pr\left(\mathcal{N}_{k-1} \cup \left(\left\{\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,k}\right\} \cap \mathcal{D}_{k-1}\right)\right) \\ &\leq \Pr\left(\mathcal{N}_{k-1}\right) + \Pr\left(\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,k} | \mathcal{D}_{k-1}\right). \end{aligned} \tag{S1.44}$$

Conditioning on $\mathcal{D}_{k-1}$ implies that the pre-selected variables $\boldsymbol{Z}_{(k-1)}$ are either signals or pseudo signals for stage k. By Assumption 11 again, the net effect of the noise variable X_l on Y with $\boldsymbol{Z}_{(k-1)}$ as pre-selected variables satisfies $\boldsymbol{\theta}_{l,\boldsymbol{Z}_{(k-1)}} \lesssim \log(m_n)^{1/2} m_n^{1/2} n^{-1/2}$. Applying the first part of Proposition 3.1 to the second term in (S1.44) yields

$$\begin{aligned}\Pr\left(\cup_{l=p^*+p^{**}+1}^{p_n} \mathcal{B}_{l,k} | \mathcal{D}_{k-1}\right) &\leq \sum_{l=p^*+p^{**}+1}^{p_n} \Pr\left(\mathcal{B}_{l,k} | \mathcal{D}_{k-1}\right) \\ &\leq \sum_{l=p^*+p^{**}+1}^{p_n} \left(n^{-M} + C_1 \exp\left(-C_2 n^{C_3}\right)\right) \\ &\leq n^{-M_1} + C_7 \exp\left(-C_5 n^{C_6}\right). \end{aligned} \tag{S1.45}$$

Combining (S1.43)–(S1.45), we obtain

$$\Pr\left(\mathcal{N}_k\right) \leq kn^{-M_1} + kC_7 \exp\left(-C_5 n^{C_6}\right),$$

and

$$\Pr\left(\mathcal{D}_k\right) = 1 - \Pr\left(\mathcal{N}_k\right) \geq 1 - kn^{-M_1} - kC_7 \exp\left(-C_5 n^{C_6}\right).$$

Apparently, when k is fixed or divergent to infinity at a rate no faster than n^a for some $a > 0$, we can write

$$\Pr\left(\mathcal{D}_k\right) \geq 1 - n^{-M_2} - C_7 \exp\left(-C_8 n^{C_9}\right) \text{ for any } k \leq n^a,$$

where M_2 is a large positive constant, $M_2 \leq M_1 - a$. ■

S2 Additional Simulation and Application Results

In this section we present some additional results for the simulation and application.

S2.1 Additional Simulation Results

In this subsection we report the simulation results for DGPs 2 - 5 and 7 - 10 in Tables S1–S8. The description of these DGPs and the summary of the simulation results are given in Sections 4.1 and 4.3 of the paper, respectively.

In order to investigate the finite sample performance of POST-OCMT procedure compared to that of OCMT procedure under linear model, we consider the following DGP.

DGP 11. Four signals, many pseudo-signals, and one hidden signal. The covariates are generated as DGP 6:

$$
\begin{aligned}
X_j &= W_j,\ j = 1, 2, \\
X_j &= \frac{W_j + U_1}{2},\ j = 3, 4, \\
X_5 &= U_1, \\
X_j &= \frac{4X_1 + (j-5)\,W_{j-1}}{j-1},\ j = 6, 10, 14, 18, \ldots, \\
X_j &= \frac{4X_2 + (j-5)\,W_{j-1}}{j-1},\ j = 7, 11, 15, 19, \ldots, \\
X_j &= \frac{4X_3 + (j-5)\,W_{j-1}}{j-1},\ j = 8, 12, 16, 20, \ldots,\ \text{and} \\
X_j &= \frac{4X_4 + (j-5)\,W_{j-1}}{j-1},\ j = 9, 13, 17, 21, \ldots
\end{aligned}
$$

where W_j, $j = 1, \ldots, p_n - 1$, and U_1 are independent draws from $U(0, 1)$. The dependent variable Y is generated as

$$
Y = 3(X_1 - 0.5) + 3(X_2 - 0.5) + 6(X_3 - 0.25) + 6\,(X_4 - 0.25) - 3X_5 + \varepsilon
$$

where ε are independent draws from $N(0, 1)$. Then, $p^* = 5$ with one hidden signal (X_5). In this DGP, we only report the results of our procedure (OCMT and Post-OCMT). The simulation results for DGP 11 are reported in Table S9. Clearly, from this table, we can see that POST-OCMT works very well in picking the true signals including the hidden one and has good out-sample forecasting performance, due to the fact that it eliminates the impact of pseudo-signals.

Table S1: DGP 2

Panel 1: $n = 200$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.1190	0.9670	0.0026	0.0452	0.7620	-	1.0795
OCMT	5.2950	0.9862	0.0141	0.2061	0.0800	1.1330	1.1899
AGLASSO	4.0710	0.9537	0.0027	0.0469	0.7490	-	1.0872
BAGGING	-	-	-	-	-	-	1.4335
RF	-	-	-	-	-	-	1.4581
Panel 2: $n = 200$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.1180	0.9690	0.0012	0.0438	0.7640	-	1.0809
OCMT	5.3560	0.9908	0.0071	0.2107	0.0740	1.1510	1.2064
AGLASSO	4.0350	0.9447	0.0013	0.0468	0.7470	-	1.0911
BAGGING	-	-	-	-	-	-	1.4621
RF	-	-	-	-	-	-	1.4914
Panel 3: $n = 200$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.1200	0.9617	0.0003	0.0494	0.7420	-	1.0825
OCMT	5.3350	0.9828	0.0014	0.2134	0.0710	1.1910	1.2387
AGLASSO	3.8194	0.8898	0.0003	0.0471	0.6863	-	1.1267
BAGGING	-	-	-	-	-	-	1.5210
RF	-	-	-	-	-	-	1.5591
Panel 4: $n = 400$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0490	0.9900	0.0009	0.0161	0.9130	-	1.0483
OCMT	5.1380	0.9942	0.0121	0.1802	0.1380	1.1000	1.1152
AGLASSO	4.0450	0.9892	0.0009	0.0161	0.9120	-	1.0468
BAGGING	-	-	-	-	-	-	1.3167
RF	-	-	-	-	-	-	1.3406
Panel 5: $n = 400$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0500	0.9848	0.0006	0.0204	0.8930	-	1.0495
OCMT	5.2130	0.9912	0.0064	0.1926	0.1040	1.0990	1.1244
AGLASSO	4.0540	0.9888	0.0005	0.0180	0.9020	-	1.0530
BAGGING	-	-	-	-	-	-	1.3475
RF	-	-	-	-	-	-	1.3752
Panel 6: $n = 400$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0480	0.9878	0.0001	0.0177	0.9050	-	1.0535
OCMT	5.1990	0.9948	0.0012	0.1875	0.1290	1.0980	1.1286
AGLASSO	4.0889	0.9874	0.0001	0.0246	0.8666	-	1.0530
BAGGING	-	-	-	-	-	-	1.4060
RF	-	-	-	-	-	-	1.4371

Table S2: DGP 3

Panel 1: $n = 200$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.4790	0.8678	0.0015	0.0202	0.4730	-	1.1418
OCMT	4.6100	0.8678	0.0028	0.0376	0.3940	1.5800	1.2230
AGLASSO	3.5440	0.6950	0.0007	0.0110	0.0100	-	1.2593
BAGGING	-	-	-	-	-	-	1.4274
RF	-	-	-	-	-	-	1.4668
Panel 2: $n = 200$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.3010	0.8284	0.0008	0.0235	0.3490	-	1.1612
OCMT	4.4600	0.8284	0.0016	0.0443	0.2710	1.4590	1.2429
AGLASSO	3.2990	0.6442	0.0004	0.0128	0.0030	-	1.2679
BAGGING	-	-	-	-	-	-	1.4584
RF	-	-	-	-	-	-	1.4971
Panel 3: $n = 200$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	3.4590	0.6730	0.0001	0.0170	0.0680	-	1.2349
OCMT	3.4900	0.6730	0.0001	0.0215	0.0590	1.0930	1.2481
AGLASSO	2.6024	0.5105	0.0001	0.0087	0.0000	-	1.3464
BAGGING	-	-	-	-	-	-	1.5109
RF	-	-	-	-	-	-	1.5544
Panel 4: $n = 400$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	5.0050	0.9990	0.0001	0.0014	0.9860	-	1.0446
OCMT	5.0330	0.9990	0.0004	0.0052	0.9650	1.9950	1.0574
AGLASSO	4.3790	0.8642	0.0006	0.0084	0.2800	-	1.1136
BAGGING	-	-	-	-	-	-	1.3272
RF	-	-	-	-	-	-	1.3683
Panel 5: $n = 400$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	5.0020	0.9972	0.0001	0.0022	0.9750	-	1.0460
OCMT	5.0460	0.9972	0.0003	0.0077	0.9470	1.9960	1.0643
AGLASSO	4.2520	0.8408	0.0002	0.0072	0.1850	-	1.1305
BAGGING	-	-	-	-	-	-	1.3574
RF	-	-	-	-	-	-	1.4001
Panel 6: $n = 400$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.8820	0.9684	0.0000	0.0057	0.8550	-	1.0632
OCMT	4.9460	0.9684	0.0001	0.0136	0.8180	1.8920	1.0924
AGLASSO	4.0764	0.8044	0.0000	0.0087	0.0255	-	1.1478
BAGGING	-	-	-	-	-	-	1.4128
RF	-	-	-	-	-	-	1.4531

Table S3: DGP 4

Panel 1: $n = 200$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.3760	0.7934	0.0043	0.0683	0.3210	-	1.1735
OCMT	5.9320	0.8120	0.0195	0.2552	0.0000	1.5270	1.5330
AGLASSO	3.1090	0.5736	0.0025	0.0477	0.0050	-	1.3469
BAGGING	-	-	-	-	-	-	1.4197
RF	-	-	-	-	-	-	1.4362
Panel 2: $n = 200$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.1070	0.7468	0.0019	0.0667	0.2580	-	1.1956
OCMT	5.6420	0.7614	0.0094	0.2617	0.0010	1.4090	1.4935
AGLASSO	2.7500	0.5052	0.0011	0.0457	0.0020	-	1.3299
BAGGING	-	-	-	-	-	-	1.4499
RF	-	-	-	-	-	-	1.4663
Panel 3: $n = 200$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	3.3360	0.5968	0.0004	0.0745	0.0470	-	1.2652
OCMT	4.5670	0.6164	0.0015	0.2472	0.0000	1.0850	1.3430
AGLASSO	2.0833	0.3773	0.0002	0.0412	0.0000	-	1.3978
BAGGING	-	-	-	-	-	-	1.5041
RF	-	-	-	-	-	-	1.5251
Panel 4: $n = 400$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	5.1300	0.9946	0.0016	0.0227	0.8440	-	1.0472
OCMT	7.0280	0.9984	0.0212	0.2531	0.0000	1.9940	1.5916
AGLASSO	4.4680	0.8426	0.0027	0.0397	0.1610	-	1.1553
BAGGING	-	-	-	-	-	-	1.3231
RF	-	-	-	-	-	-	1.3455
Panel 5: $n = 400$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	5.0930	0.9924	0.0007	0.0190	0.8620	-	1.0476
OCMT	7.0330	0.9960	0.0105	0.2545	0.0000	1.9930	1.6081
AGLASSO	4.3460	0.8278	0.0011	0.0326	0.1050	-	1.1343
BAGGING	-	-	-	-	-	-	1.3522
RF	-	-	-	-	-	-	1.3764
Panel 6: $n = 400$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.9700	0.9598	0.0002	0.0255	0.7410	-	1.0660
OCMT	6.8910	0.9656	0.0021	0.2609	0.0000	1.8850	1.5604
AGLASSO	4.1797	0.7901	0.0002	0.0383	0.0208	-	1.1478
BAGGING	-	-	-	-	-	-	1.4073
RF	-	-	-	-	-	-	1.4309

Table S4: DGP 5

	NV	TPR	FPR	FDR	CS	STEP	RMSFE
Panel 1: $n = 200$, $p = 100$							
POST OCMT	3.4700	0.8465	0.0009	0.0141	0.3580	-	1.1198
OCMT	3.5210	0.8465	0.0014	0.0223	0.3120	1.4090	1.3958
AGLASSO	2.8930	0.7127	0.0004	0.0073	0.0800	-	1.1856
BAGGING	-	-	-	-	-	-	1.2298
RF	-	-	-	-	-	-	1.2677
Panel 2: $n = 200$, $p = 200$							
POST OCMT	3.4040	0.8283	0.0005	0.0154	0.2930	-	1.1235
OCMT	3.4630	0.8283	0.0008	0.0249	0.2430	1.3510	1.3660
AGLASSO	2.7600	0.6793	0.0002	0.0083	0.0490	-	1.1822
BAGGING	-	-	-	-	-	-	1.2512
RF	-	-	-	-	-	-	1.2906
Panel 3: $n = 200$, $p = 1000$							
POST OCMT	2.9990	0.7380	0.0000	0.0089	0.0500	-	1.1515
OCMT	3.0080	0.7380	0.0001	0.0103	0.0440	1.0650	1.1970
AGLASSO	2.4560	0.6030	0.0000	0.0087	0.0150	-	1.2163
BAGGING	-	-	-	-	-	-	1.2904
RF	-	-	-	-	-	-	1.3293
Panel 4: $n = 400$, $p = 100$							
POST OCMT	3.9800	0.9908	0.0002	0.0028	0.9460	-	1.0420
OCMT	4.0300	0.9908	0.0007	0.0110	0.8990	1.9020	1.6559
AGLASSO	3.8250	0.9460	0.0004	0.0068	0.7470	-	1.0555
BAGGING	-	-	-	-	-	-	1.1604
RF	-	-	-	-	-	-	1.1922
Panel 5: $n = 400$, $p = 200$							
POST OCMT	3.9670	0.9850	0.0001	0.0044	0.9150	-	1.0492
OCMT	4.0260	0.9850	0.0004	0.0139	0.8630	1.9210	1.6745
AGLASSO	3.7640	0.9300	0.0002	0.0075	0.6820	-	1.0676
BAGGING	-	-	-	-	-	-	1.1751
RF	-	-	-	-	-	-	1.2147
Panel 6: $n = 400$, $p = 1000$							
POST OCMT	3.7550	0.9300	0.0000	0.0060	0.6900	-	1.0663
OCMT	3.8200	0.9300	0.0001	0.0164	0.6350	1.7200	1.5583
AGLASSO	3.5100	0.8675	0.0000	0.0069	0.4500	-	1.0865
BAGGING	-	-	-	-	-	-	1.2045
RF	-	-	-	-	-	-	1.2520

Table S5: DGP 7

Panel 1: $n = 200$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0030	1.0000	0.0000	0.0005	0.9970	-	1.0793
OCMT	4.0080	1.0000	0.0001	0.0013	0.9920	1.0240	1.0922
AGLASSO	4.0470	0.9970	0.0006	0.0097	0.9400	-	1.0858
BAGGING	-	-	-	-	-	-	1.4393
RF	-	-	-	-	-	-	1.4824
Panel 2: $n = 200$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0070	0.9992	0.0001	0.0017	0.9870	-	1.0833
OCMT	4.0120	0.9992	0.0001	0.0025	0.9820	1.0450	1.1071
AGLASSO	4.0500	0.9932	0.0004	0.0125	0.9240	-	1.0906
BAGGING	-	-	-	-	-	-	1.4758
RF	-	-	-	-	-	-	1.5246
Panel 3: $n = 200$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0050	0.9935	0.0000	0.0052	0.9500	-	1.0873
OCMT	4.0300	0.9935	0.0001	0.0088	0.9360	1.0910	1.1260
AGLASSO	3.9690	0.9643	0.0001	0.0180	0.8630	-	1.1108
BAGGING	-	-	-	-	-	-	1.5348
RF	-	-	-	-	-	-	1.5919
Panel 4: $n = 400$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0010	1.0000	0.0000	0.0002	0.9990	-	1.0601
OCMT	4.0010	1.0000	0.0000	0.0002	0.9990	1.0000	1.0601
AGLASSO	4.0210	1.0000	0.0002	0.0035	0.9790	-	1.0608
BAGGING	-	-	-	-	-	-	1.3339
RF	-	-	-	-	-	-	1.3755
Panel 5: $n = 400$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0000	1.0000	0.0000	0.0000	1.0000	-	1.0575
OCMT	4.0000	1.0000	0.0000	0.0000	1.0000	1.0000	1.0575
AGLASSO	4.0230	1.0000	0.0001	0.0038	0.9770	-	1.0585
BAGGING	-	-	-	-	-	-	1.3646
RF	-	-	-	-	-	-	1.4083
Panel 6: $n = 400$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0020	1.0000	0.0000	0.0003	0.9980	-	1.0582
OCMT	4.0020	1.0000	0.0000	0.0003	0.9980	1.0000	1.0582
AGLASSO	4.0530	1.0000	0.0001	0.0086	0.9520	-	1.0601
BAGGING	-	-	-	-	-	-	1.4165
RF	-	-	-	-	-	-	1.4619

Table S6: DGP 8

Panel 1: $n = 200$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	3.5670	0.6606	0.0027	0.0527	0.1420	-	1.2378
OCMT	4.7090	0.6784	0.0137	0.2056	0.0000	1.2270	1.3827
AGLASSO	2.9240	0.5442	0.0021	0.0403	0.0010	-	1.3769
BAGGING	-	-	-	-	-	-	1.4590
RF	-	-	-	-	-	-	1.4697
Panel 2: $n = 200$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	3.4800	0.6370	0.0015	0.0605	0.1180	-	1.2535
OCMT	4.5760	0.6556	0.0066	0.2076	0.0000	1.1760	1.3681
AGLASSO	2.6250	0.4824	0.0011	0.0425	0.0010	-	1.3807
BAGGING	-	-	-	-	-	-	1.4883
RF	-	-	-	-	-	-	1.4998
Panel 3: $n = 200$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	3.2590	0.5874	0.0003	0.0685	0.0420	-	1.2762
OCMT	4.2030	0.6082	0.0012	0.1994	0.0000	1.0790	1.3341
AGLASSO	1.9940	0.3668	0.0002	0.0336	0.0000	-	1.4378
BAGGING	-	-	-	-	-	-	1.5429
RF	-	-	-	-	-	-	1.5574
Panel 4: $n = 400$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.8870	0.9474	0.0016	0.0230	0.7160	-	1.0743
OCMT	7.4610	0.9534	0.0281	0.3053	0.0000	1.8440	1.5024
AGLASSO	4.3690	0.8278	0.0024	0.0363	0.1030	-	1.1364
BAGGING	-	-	-	-	-	-	1.3586
RF	-	-	-	-	-	-	1.3840
Panel 5: $n = 400$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.8820	0.9432	0.0008	0.0249	0.6890	-	1.0768
OCMT	7.4170	0.9496	0.0136	0.3032	0.0000	1.8290	1.4988
AGLASSO	4.2570	0.8090	0.0011	0.0351	0.0550	-	1.1455
BAGGING	-	-	-	-	-	-	1.3882
RF	-	-	-	-	-	-	1.4130
Panel 6: $n = 400$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.8420	0.9338	0.0002	0.0262	0.6490	-	1.0825
OCMT	7.3220	0.9382	0.0026	0.3020	0.0000	1.7920	1.4947
AGLASSO	4.1570	0.7908	0.0002	0.0346	0.0120	-	1.1550
BAGGING	-	-	-	-	-	-	1.4532
RF	-	-	-	-	-	-	1.4732

Table S7: DGP 9

Panel 1: $n = 200$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.2770	0.9945	0.0005	0.0031	0.7380	-	1.0422
OCMT	4.2950	0.9948	0.0005	0.0033	0.7300	1.1000	1.0436
AGLASSO	5.0510	1.0000	0.0014	0.0109	0.4800	-	1.0537
BAGGING	-	-	-	-	-	-	1.3471
RF	-	-	-	-	-	-	1.3909
Panel 2: $n = 200$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.5000	0.9870	0.0004	0.0028	0.5780	-	1.0585
OCMT	4.5310	0.9870	0.0004	0.0030	0.5650	1.1460	1.0585
AGLASSO	5.7540	0.9995	0.0011	0.0090	0.3420	-	1.0697
BAGGING	-	-	-	-	-	-	1.3838
RF	-	-	-	-	-	-	1.4351
Panel 3: $n = 200$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.8200	0.9165	0.0002	0.0012	0.2200	-	1.1207
OCMT	4.8670	0.9170	0.0002	0.0012	0.2120	1.1410	1.1236
AGLASSO	8.7040	0.9890	0.0004	0.0048	0.1060	-	1.1143
BAGGING	-	-	-	-	-	-	1.4450
RF	-	-	-	-	-	-	1.5038
Panel 4: $n = 400$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0110	1.0000	0.0000	0.0001	0.9890	-	1.0220
OCMT	4.0110	1.0000	0.0000	0.0001	0.9890	1.0080	1.0220
AGLASSO	4.6660	1.0000	0.0010	0.0069	0.6240	-	1.0288
BAGGING	-	-	-	-	-	-	1.2348
RF	-	-	-	-	-	-	1.2713
Panel 5: $n = 400$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0150	1.0000	0.0000	0.0001	0.9860	-	1.0237
OCMT	4.0150	1.0000	0.0000	0.0001	0.9860	1.0080	1.0237
AGLASSO	5.2000	1.0000	0.0008	0.0061	0.4460	-	1.0351
BAGGING	-	-	-	-	-	-	1.2620
RF	-	-	-	-	-	-	1.3032
Panel 6: $n = 400$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0550	1.0000	0.0000	0.0001	0.9490	-	1.0225
OCMT	4.0570	1.0000	0.0000	0.0001	0.9480	1.0360	1.0225
AGLASSO	7.7000	1.0000	0.0004	0.0037	0.1270	-	1.0551
BAGGING	-	-	-	-	-	-	1.3132
RF	-	-	-	-	-	-	1.3576

Table S8: DGP 10

Panel 1: $n = 200$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0240	0.9668	0.0016	0.0260	0.7450	-	1.0329
OCMT	4.0250	0.9668	0.0016	0.0261	0.7440	1.8780	1.0329
AGLASSO	5.0160	0.9942	0.0108	0.1413	0.4370	-	1.0375
BAGGING	-	-	-	-	-	-	1.2856
RF	-	-	-	-	-	-	1.3107
Panel 2: $n = 200$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0700	0.9535	0.0013	0.0403	0.6560	-	1.0421
OCMT	4.0740	0.9535	0.0013	0.0409	0.6540	1.8320	1.0421
AGLASSO	5.7520	0.9915	0.0091	0.2184	0.2760	-	1.0519
BAGGING	-	-	-	-	-	-	1.3212
RF	-	-	-	-	-	-	1.3391
Panel 3: $n = 200$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0410	0.8625	0.0006	0.0899	0.2740	-	1.0910
OCMT	4.0490	0.8628	0.0006	0.0908	0.2740	1.5650	1.0922
AGLASSO	7.7110	0.9473	0.0039	0.3834	0.0530	-	1.1009
BAGGING	-	-	-	-	-	-	1.3701
RF	-	-	-	-	-	-	1.3878
Panel 4: $n = 400$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0060	0.9998	0.0001	0.0012	0.9920	-	1.0068
OCMT	4.0060	0.9998	0.0001	0.0012	0.9920	1.9870	1.0068
AGLASSO	4.6880	1.0000	0.0072	0.0972	0.5930	-	1.0135
BAGGING	-	-	-	-	-	-	1.1703
RF	-	-	-	-	-	-	1.2185
Panel 5: $n = 400$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0050	0.9998	0.0000	0.0010	0.9930	-	1.0081
OCMT	4.0050	0.9998	0.0000	0.0010	0.9930	1.9950	1.0081
AGLASSO	5.2130	1.0000	0.0062	0.1571	0.4460	-	1.0195
BAGGING	-	-	-	-	-	-	1.1953
RF	-	-	-	-	-	-	1.2466
Panel 6: $n = 400$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.0170	0.9985	0.0000	0.0038	0.9710	-	1.0089
OCMT	4.0170	0.9985	0.0000	0.0038	0.9710	1.9920	1.0089
AGLASSO	7.2690	0.9998	0.0033	0.3310	0.1700	-	1.0373
BAGGING	-	-	-	-	-	-	1.2492
RF	-	-	-	-	-	-	1.2888

Table S9: DGP 11

Panel 1: $n = 200$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.9320	0.9290	0.0445	0.0030	0.5490	-	1.0928
OCMT	6.4460	0.9404	0.2162	0.0182	0.0560	1.5020	1.5047
Panel 2: $n = 200$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.9770	0.9328	0.0486	0.0016	0.5720	-	1.0888
OCMT	6.5490	0.9444	0.2266	0.0093	0.0480	1.4550	1.4812
Panel 3: $n = 200$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	4.9270	0.9250	0.0469	0.0003	0.5520	-	1.0960
OCMT	6.5430	0.9382	0.2280	0.0019	0.0500	1.4590	1.4843
Panel 4: $n = 400$, $p = 100$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	5.0450	0.9870	0.0168	0.0011	0.8780	-	1.0298
OCMT	6.6390	0.9894	0.2077	0.0176	0.0940	1.3610	1.3662
Panel 5: $n = 400$, $p = 200$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	5.0350	0.9860	0.0159	0.0005	0.8660	-	1.0330
OCMT	6.6150	0.9866	0.2056	0.0086	0.1050	1.3640	1.3687
Panel 6: $n = 400$, $p = 1000$							
	NV	TPR	FPR	FDR	CS	STEP	RMSFE
POST OCMT	5.0320	0.9872	0.0145	0.0001	0.8810	-	1.0360
OCMT	6.6970	0.9886	0.2061	0.0018	0.0890	1.3820	1.3857

S2.2 Additional Application Results

In this subsection, we first provide the description of the variables used in the application and some summary statistics. Then we report the frequencies of variables selected.

Table S10 gives the description of the dependent variable and covariates used in our application. Table S11 reports the summary statistics for the variables. Table S12 reports the frequencies of covariates selected by the methods under investigation. For those variables that are not listed in Table S12, they are not selected by any of these methods. Table S12 suggests that among the 78 covariates in the application, three of them, namely, G102, G133, and G137, are all frequently selected by the multiple-stage OCMT, post-OCMT, group Lasso and adaptive group Lasso.

Table S10: Definitions of Variables

Panel A: Dependent Variable	
$G136$	Remittance to his/her original home in rural areas in 2007

Panel B: Continuous Independent Variables	
$A05$	Age
$A08$	Rank of Siblings (1: the oldest, 2: the second oldest, and so on so forth)
$A25$	Height (in cm)
$A26$	Weight (in kg)
$A37$	Cost of medical insurance in 2007 (in RMB)
$A39$	Total medical cost in 2007 (in RMB)
$A40$	Out-of-Pocket medical cost in 2007 (in RMB)
$B103$	Years of eduction
$C103$	Year when starting current job
$C111$	Number of employees in the company he/she is working (ranked from 1 (smallest) to 9 (largest))
$C112$	Average number of working hours per week
$C105$	Date when working in current occupation
$C117$	Average income in current occupation
$C126$	Food compensation per month from current job (in RMB)
$C165$	Date when start to work in urban areas
$C171$	Time spent to find the first job (in days)
$C177$	Average number of working hours per week for the fist job
$C178$	The salary of the first month for the first job (in RMB)
$C179$	The salary of the last month for the first job (in RMB)
$C183$	Number of months for doing the first job
$C186$	Number of cities in which he/she worked
$C188$	Average month salary if he/she worked in his/her hometown

Continued on next page

Table S10 – continued from previous page

$E4\_11$	Number of people he/she contacted during the last spring festival
$E4\_12$	Number of relatives he/she contacted during the last spring festival
$E4\_13$	Number of close friends he/she contacted during the last spring festival
$E4\_14$	Number of people in major cities he/she contacted during the last spring festival
$E4\_16$	Number of people who helped him/her in the last 12 months
$G119$	Average cost of food per month including dining in restaurants (in RMB)
$G120$	Average cost on clothes per month (in RMB)
$G121$	Average cost of accommodation including utilities per month (in RMB)
$G122$	Total living cost in 2007 (in RMB)
$G123$	Total consumption of durable goods in 2007 (in RMB)
$G124$	Total consumption of non-durable goods or services (e.g., dishes, beauty, hair cutting) in 2007 (in RMB)
$G125$	Total consumption of medication, nutrition supplements, and physical therapies in 2007 (in RMB)
$G126$	Cost of transportation in 2007 (in RMB)
$G127$	Cost of phone services and postage in 2007 (in RMB)
$G128$	Cost of entertainment in 2007 (in RMB)
$G132$	Other consumption in 2007 (in RMB)
$G141$	Cost per month for the minimum living standard in current city (in RMB)
$H219\_2$	Value of his/her cell phones (in RMB)
$I101$	Current total living area (in square meters)
$I112$	Current rent per month (in RMB)
$J108$	Average cost of hiring a handy man per day in his/her hometown (in RMB)
$J109$	Percentage of workforce working in cities in his/her country
$J112$	Value of houses owned in his/her hometown (in RMB)
$J113$	Acres of land owned in his/her hometown
$G102$	Monthly income (in RMB)

Panel C: Continuous Independent Variables but Treated as Dummy Variables

$C130$	Compensation per month for accommodation from the current job (in RMB)
$E4\_15$	Number of people in possession of city Hukou he/she contacted during the last spring festival
$G129$	Cost of education for children excluding left-behind children in 2007 (in RMB)
$G130$	Cost of all non-saving insurances in 2007 (in RMB)
$G131$	Other consumption in 2007 (in RMB)
$G133$	Gifts to others including his/her parents in 2007 (in RMB)
$G134$	Investments in stocks or bonds in 2007 (in RMB)

Continued on next page

Table S10 – continued from previous page

$G135$	Cost of building new houses or renovation in his/her hometown in 2007 (in RMB)
$G137$	Education cost for left-behind children in 2007 (in RMB)
$G138$	Fines or interests charged in 2007 (in RMB)
$G139$	Cost of loans in 2007 (in RMB)
$B110$	Total marks of his/her college entrance exam (0 if not attended)
$B119$	Cost of last job training (in RMB)

Panel D: Discrete Variables that Treated as Dummy Variables

$A01$	Number of months working outside of his/her hometown in 2007
$A10$	Number of children
$A27$	Health condition (excellent to bad ranking from 1 to 5)
$H219\_1$	Number of cell phones owned
$I102$	Number of people currently living together

Panel E: Dummy Variables

$A04$	Gender (0: female 1: male)
$A09$	Marital Status (0: not married 1:married)
$A15$	Status of Hukou (0: rural 1: city)
$A21$	Ownership of the insurance for unemployment (0: no 1: yes)
$A22$	Ownership of the insurance for age care (0: no 1: yes)
$A23$	Ownership of the insurance for injuries during work (0: no 1: yes)
$A24$	Ownership of the programme of the deposit for house purchases (0: no 1: yes)
$A28$	Disable or not (0: no 1:yes)
$B108$	Whether he/she attended the college entrance exam or not (0: no 1:yes)
$C127$	Whether employer provides housing or not (0: no 1:yes)
$C184$	Whether he/she ever lived in his/her hometown for over 3 months continuously after working in cities (0: no 1:yes)
Dongguan	Whether he/she works in Dong Guan city (0: no 1:yes)
Shenzhen	Whether he/she works in Shen Zhen city (0: no 1:yes)

Table S11: Summary Statistics

Variable	MEAN	STD	MIN	LQ	MD	UQ	MAX
Panel A: Dependent Variable							
$G136$	3374	4576	0	0	2000	5000	60000
Panel B: Continuous Independent Variables							

Continued on next page

Table S11 – continued from previous page

Variable	MEAN	STD	MIN	LQ	MD	UQ	MAX
$A05$	30	9	17	23	27	36	66
$A08$	2	1	1	1	2	3	8
$A25$	166	7	147	160	168	171	186
$A26$	59	9	36	52	60	65	82
$A37$	109	413	0	0	10	10	5000
$A39$	289	746	0	0	50	270	8000
$A40$	261	676	0	0	50	200	8000
$B103$	10	2	2	9	9	12	20
$C103$	2006	3	1989	2005	2006	2007	2008
$C111$	6	2	1	4	6	7	9
$C112$	57	14	34	48	56	70	126
$C105$	200516	314	198910	200404	200606	200709	200811
$C117$	1620	795	500	1165	1500	1900	10000
$C126$	102	112	0	0	95	160	450
$C165$	199896	650	197607	199555	200012	200404	200803
$C171$	14	26	0	2	7	15	365
$C177$	58	19	0	48	60	70	114
$C178$	724	423	0	500	700	900	5000
$C179$	874	516	0	550	800	1100	5000
$C183$	18	24	1	6	12	24	188
$C186$	2	5	0	1	2	3	90
$C188$	858	853	0	600	800	1000	10000
$E4\_11$	38	55	0	15	29	40	800
$E4\_12$	15	14	0	6	10	20	120
$E4\_13$	24	50	0	5	11	20	700
$E4\_14$	18	34	0	3	10	20	400
$E4\_16$	4	11	0	0	2	3	200
$G119$	520	324	0	300	500	700	3000
$G120$	125	153	0	0	100	200	1500
$G121$	207	302	0	0	100	300	2000
$G122$	14630	9131	0	8795	12490	18425	75300
$G123$	726	1385	0	0	200	1000	10000
$G124$	876	2080	0	300	500	1000	37000
$G125$	380	850	0	0	100	400	8500
$G126$	701	843	0	200	500	1000	7200
$G127$	971	989	0	515	900	1200	17400

Continued on next page

Table S11 – continued from previous page

Variable	MEAN	STD	MIN	LQ	MD	UQ	MAX
$G128$	302	664	0	0	100	300	5000
$G132$	4813	6800	0	1000	3000	6000	74400
$G141$	1031	748	50	500	900	1400	8700
$H219\_2$	907	750	10	300	800	1200	5000
$I101$	28	27	3	15	20	30	325
$I112$	161	239	0	0	50	250	1800
$J108$	38	13	0	30	40	50	80
$J109$	59	17	0	50	60	70	90
$J112$	52528	71220	0	10000	30000	80000	700000
$J113$	1	2	0	1	1	1	20
$G102$	1756	1055	0	1100	1500	2200	11000
Panel C: Continuous Independent Variables but Treated as Dummy Variables							
$C130$	72	97	0	0	0	150	500
$E4\_15$	8	18	0	0	2	7	250
$G129$	251	2164	0	0	0	0	40400
$G130$	33	232	0	0	0	0	3000
$G131$	194	1028	0	0	0	0	18800
$G133$	422	1088	0	0	0	400	12000
$G134$	80	1193	0	0	0	0	20000
$G135$	164	2758	0	0	0	0	60000
$G137$	535	2279	0	0	0	0	27000
$G138$	149	625	0	0	0	0	7800
$G139$	1776	4464	0	0	0	1000	45000
$B110$	45	134	0	0	0	0	630
$B119$	103	633	0	0	0	0	10000
Panel D: Discrete Variables that Treated as Dummy Variables							
$A01$	11	2	1	12	12	12	12
$A10$	0.665	0.838	0	0	0	1	4
$A27$	1.749	0.770	1	1	2	2	5
$H219\_1$	1.214	0.563	0	1	1	1	4
$I102$	4.278	3.268	1	2	3	6	20
Panel E: Dummy Variables							
$A04$	0.691	0.462	0	0	1	1	1

Continued on next page

Table S11 – continued from previous page

Variable	MEAN	STD	MIN	LQ	MD	UQ	MAX
$A09$	0.486	0.500	0	0	0	1	1
$A15$	0.014	0.119	0	0	0	0	1
$A21$	0.189	0.392	0	0	0	0	1
$A22$	0.348	0.477	0	0	0	1	1
$A23$	0.356	0.479	0	0	0	1	1
$A24$	0.029	0.167	0	0	0	0	1
$A28$	0.021	0.142	0	0	0	0	1
$B108$	0.132	0.338	0	0	0	0	1
$C127$	0.698	0.460	0	0	1	1	1
$C184$	0.220	0.415	0	0	0	0	1
Dongguan	0.317	0.466	0	0	0	1	1
Shenzhen	0.374	0.484	0	0	0	1	1

Note: MEAN = sample mean, STD = standard deviation, MIN = minimum, LQ = 25 percentile, MD = median, UQ = 75 percentile, MAX = maximum.

Table S12: Frequencies of Variables Selected

	One Stage	Multiple Stage	Post-OCMT	Group LASSO	Adaptive GLASSO
$G102$	100%	100%	100%	100%	100%
$G133$	0%	98%	91%	100%	100%
$G137$	12%	98%	91%	85%	85%
$A09$	0%	1%	0%	0%	0%
$C111$	0%	0%	0%	2%	0%
$C112$	0%	0%	0%	58%	55%
$C126$	0%	0%	0%	11%	10%
$C188$	0%	0%	0%	17%	15%
$A04$	0%	0%	0%	4%	2%
$C127$	0%	0%	0%	3%	2%
$C127$	0%	0%	0%	3%	2%
$I102$	0%	0%	0%	1%	0%
$G134$	0%	0%	0%	4%	3%
$G135$	0%	0%	0%	22%	20%
$G138$	0%	8%	6%	0%	0%
$G139$	0%	0%	0%	4%	0%

References

Bernstein, D. S. (2005): *Matrix Mathematics: Theory, Facts, and Formulas*, Princeton University Press.

Stone, C. J. (1985): "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, Vol 13, No 2, 689-705.