TESTING-BASED FORWARD MODEL SELECTION

DAMIAN KOZBUR

ETH ZÜRICH

D-MATH, SEMINAR FÜR STATISTIK
RÄMISTRASSE 101, 8092 ZÜRICH

D-GESS, CENTER FOR LAW AND ECONOMICS
HALDENEGGSTEIG 4, 8092 ZÜRICH

Abstract. This work introduces a theoretical foundation for a procedure called 'testing-based forward model selection' in regression problems. Forward selection is a general term refering to a model selection procedure which inductively selects covariates that add predictive power into a working statistical model. This paper considers the use of testing procedures, derived from traditional statistical hypothesis testing, as a criterion for deciding which variable to include next and when to stop including variables. Probabilistic bounds for prediction error and number of selected covariates are proved for the proposed procedure. The general result is illustrated by an example with heteroskedastic data where Huber-Eicker-White standard errors are used to construct tests. The performance of the testing-based forward selection is compared to Lasso and Post-Lasso in simulation studies. Finally, the use of testing-based forward selection is illustrated with an application to estimating the effects of institution quality on aggregate economic output.

1. Introduction

This paper considers model selection using an algorithm called Testing-Based Forward Selection. In general, forward selection algorithms are simple and common model selection procedures that inductively select covariates which substantially increase predictive accuracy into a working statistical model until a stopping criterion is met. A leading example is in the linear regression model, where forward selection steps choose the variable that

---

Date: First version: Jan 2015. This version is of November 15, 2015. I gratefully acknowledge helpful discussion with Christian Hansen, Tim Conley, Attendants at the ETH Zürich Seminar für Statistik Research Seminar, Attendants at the Center for Law and Economics Internal Seminar, as well as financial support of the ETH Fellowship program.

gives the highest increase of in-sample-R-squared above the previous working model.

In practice, deciding which covariate gives the best additional predictive power is complicated by the fact that outcomes are observed with noise or are partly idiosyncratic. For example, in linear regression, a variable associated to a positive increment of in-sample R-squared upon inclusion to a statistical model may not add any predictive power out-of-sample. Statistical hypothesis tests offer one way to determine whether a variable of interest is likely to improve out-of-sample predictions. Furthermore, in many econometric and statistical applications, the classical assumption of independent and identically distributed data is not always appropriate. One example of this is the presence of heteroskedastic disturbances. In such settings, higher R-squared resulting from inclusion of one variable relative to another need not be a signal that the first variable is a better choice. More generally, model selection procedures tailored to the classical assumptions may have inferior performance when applied to more realistic data generating processes. The availability of hypothesis tests for diverse classes of problems and settings motivates us to introduce a testing-based model selection strategy.

We are interested in application of model selection for high-dimensional data. High-dimensional data is characterized as data with a large number of covariates relative to the sample size. High-dimensional data arise through a combination of two ways; the data may be intrinsically high dimensional in that many different characteristics per observation are available; alternatively, even when the number of available variables is relatively small, researchers rarely know the exact functional form with which the variables enter the model of interest and are thus faced with a large set of potential variables formed by different ways of interacting and transforming the underlying variables.

Dealing with a high-dimensional dataset necessarily involves dimension reduction or regularization. A principal goal of research in high-dimensional statistics and econometrics is to generate predictive power that guards against false discovery and overfitting, does not erroneously equate in-sample fit to out-of-sample predictive ability, and accurately accounts for using the same data to examine many different hypotheses or models. Without dimension reduction or regularization, however, any statistical model will overfit a high dimensional dataset. In this light, we are interested in understanding the behavior of testing-based forward selection since it potentially offers a completely data-driven way to regularize high dimensional models.

In economics, models learned using formal model selection are often used in subsequent estimation steps. A prime application of model selection is for structural estimation. One example is the selection of instrumental variables for later use in a first stage regression (see [5], [22]). Another example is the selection of a conditioning set, to properly control for omitted variables bias when there are many control variables (see [9], [41], [7], [28]). In both cases, bounds about the quality of the selected model are used to derive results about the quality of post-model selection estimation and guide subsequent inference. This paper provides the first adequately tight bounds using strictly forward selection for application in causal post-estimation analysis.

Another motivation for studying forward selection algorithms is that they are potentially computationally efficient. The results proven in this paper provide further guarantees that the potential speed up in computational efficiency does not come at a high cost in terms of statistical efficiency.

There are several earlier analyses of forward selection. Previous papers do not attempt to make use of testing as a criteria for stopping. [43] gives an bounds on the performance and number of selected covariates under a β-min condition which restricts the minimum magnitude of nonzero coefficients. [46] and [39] prove performance bounds greedy algorithms under a strong irrepresentability condition, which restricts the empirical covariance matrix of the predictors. [18] prove bounds on the relative performance in population R-squared of the a forward selection based model (relative to infeasible R-squared) when the number of variables allowed for selection is fixed. In this paper, we prove probabilistic bounds on the predictive performance and number of selected covariates. We use conditions which are much weaker that those used in [46] and [39], and impose no β-min restrictions.

There are many other approaches to high dimensional estimation and regularization. An important and common approach to generic high dimensional estimation problems are the Lasso and Post-Lasso estimations. The Lasso minimizes a least squares criteria augmented with a penalty proportional to the ℓ_1 norm of the coefficient vector. This approach favors a model with good in sample prediction while still placing high value on parsimony (the structure of the objective sets many coefficients are set identically to zero). The Post-Lasso refits based on a least squares objective function on the selected model. For theoretical and simulation results about the performance of these two methods, see [19] [38], [23] [16] [3], [4], [11], [14], [13] [15], [16], [24], [26], [27], [29], [30], [32], [34], [38], [40], [42], [45], [6], [12], [6], among many more. In terms of the convergence results in this paper, ours are likely most similar to the analysis of a forward-backword model selection procedure by Tong Zhang (see [47]).

We are interested in the relative performance of testing based forward selection relative to Lasso and Post-Lasso. A potential benefit of testing-based forward selection relative to Lasso is that it is much more easily adapted to a diverse set of problems. Forward selection can be applied to virtually any problem for which there is a reliable testing procedure for determining whether any particular variable (or set of variables or new parameters) adds predictive power.

We derive bounds for forward selection which are qualitatively similar to those given by Lasso. The proofs of these facts are original and require a fundamentally different analysis than the common logic for Lasso, partly because there is no single objective function guiding the model selection process. The argument requires us to keep track of the relative sizes of the signals individual covariates carry about the outcome. We characterize the geometric relations of the covariates carrying weak signals about the outcome relative to the covariates which are strong predictors. We accomplish this without β-min conditions. The general result is illustrated by an example with heteroskedastic data where Huber-Eicker-White standard errors are used to construct tests. We provide simulation results to show relative performance to Lasso and Post-Lasso regression. We find that there are data generating processes under which forward selection outperforms Lasso regression in terms of prediction.

Finally, we illustrate the use of testing-based forward selection in an economic application. We revisit the question studied by Acemoglu, Johnson and Robinson (see [1]) of learning the effect of institution quality on aggregate economic output in a cross section of 64 countries. [1] propose an instrumental variables strategy, using early European settler mortality rates as an instrument for current quality of institutions as measured the extent of protection from expropriation. They provide an argument concluding that the effect of institutions on output can be identified using early settler mortality as an instrument, provided that geography is properly controlled for. In their baseline specification, [1] address this by including a variable equal to latitude. However, geography is a broad notion and can potentially mean many different things; for example, temperature, yearly rainfall, terrain. As a compliment to their analysis, we consider 16 different possible controls for geography. We use testing-based forward selection to choose the most relevant geographic controls. To be robust to model selection mistakes and not suffer classical problems known to be associated with pretesting, we require three model selection steps (see [8], [9]), each taking a separate application of testing-based model selection. These are: (1) We select those geographic variables predictive of output; (2) We select those geographic controls predictive of quality of institution; (3) We select those geographic controls predictive of European settler mortality. Finally, we perform standard IV estimation using the union of selected controls. Our findings about the effects of institutions on output are largely consistent with theirs when model selection is used to determine the way to control for geography. Interestingly, this provides further evidence supporting the robustness of the conclusions made in [1].

2. Framework

Consider random variables $\{y_i\}_{i=1}^n \in \mathcal{Y}^n \subset \mathbb{R}^n$ and a set of covariates $\{x_i\}_{i=1}^n \in \mathcal{X}^n$ which are jointly distributed according to a distribution P. We are interested in constructing a function

$$\widehat{f} : \mathcal{X} \to \mathcal{Y}$$

such $\{\widehat{f}(x_i)\}_{i=1}^n$ gives good predictions about $\{y_i\}_{i=1}^n$ according to an appropriate measure of loss. Consider a family of loss functions indexed by $f \in \mathbf{F}$ which in this paper will always be quadratic:

$$\ell_f : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$$

$$\ell_f(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2.$$

We will consider the following set of approximating functions to $\mathbf{F}$,

$$\mathcal{F} = \left\{ f_\theta(\cdot) = \sum_{k=1}^p \theta_k \psi_k(\cdot), \theta \in \Theta \right\},$$

and we assume that $\mathcal{F} \subset \mathbf{F}$. Common choices for $\mathcal{F}$ include orthogonal polynomials, b-splines, or simply the components of x_i themselves when $\mathcal{X} = \mathbb{R}^p$. We are interested in finding a value θ which minimizes

$$\mathcal{E}(\theta) := \mathrm{E}\ell_{f_\theta} - \inf_{f \in \mathbf{F}} \mathrm{E}\ell_f$$

where E is used to denote the expectation operator with respect to P. We procede by searching for a sparse subset $\widehat{S} \subset \{1, ..., p\}$ that assumes a small value of

$$\mathcal{E}(S) := \inf_{\mathrm{supp}(\theta) \subset S} \mathrm{E}\ell_{f_\theta} - \inf_{f \in \mathbf{F}} \mathrm{E}\ell_f,$$

estimating θ with

$$\widehat{\theta} \in \arg \min_{\mathrm{supp}(\theta) \subset \widehat{S}} \ell_{f_\theta}(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n)$$

and finally constructing

$$\widehat{f}(\cdot) = f_{\widehat{\theta}}(\cdot).$$

The goal is to select $\widehat{S}$ by a forward selection procedure which involves the use of statistical hypothesis tests. For any S define the incremental loss from the jth covariate by

$$\Delta_j \mathcal{E}(S) = \mathcal{E}(S \cup \{j\}) - \mathcal{E}(S).$$

We consider a greedy algorithm which inductively selects the jth covariate to enter a working model if $\Delta_j \mathcal{E}(S)$ is large and $\Delta_j \mathcal{E}(S) \geqslant \Delta_k \mathcal{E}(S)$ for each $k \neq j$. However, $\Delta_j \mathcal{E}(S)$ cannot be directly observed from any single realization of the data. Therefore, we make use of statistical tests to gauge the magnitude of $\Delta_j \mathcal{E}(S)$.

Consider a set of tests which will guide the forward selection process:

$$T_{jS\alpha} \in \{0,1\} \text{ associated to } H_0 : \Delta_j \mathcal{E}(S) = 0 \text{ and level } \alpha > 0.$$

We assume that the tests take a value of $T_{jS\alpha} = 1$ for large values of a test statistic W_{jS}. Therefore, large values of the random variables W_{jS} $\Delta_j \mathcal{E}(S)$ are tied to large values of $\Delta_j \mathcal{E}(S)$ in a way made precise below.

The model selection procedure is as follows. Start with an empty model (consisting of no covariates). At each step, if the current model is $\widehat{S}$, select one covariate such that $T_{j\widehat{S}\alpha} = 1$, append it to $\widehat{S}$, and continue to the next step; if no covariates have $T_{j\widehat{S}\alpha} = 1$, then terminate the model selection procedure and return the current model. If at any juncture, there are two indices j, k (or more) such that $T_{jS\alpha} = T_{kS\alpha} = 1$, the selection is made according to the larger value of W_{jS}, W_{kS}. Alternatively, we could have devised additional tests $T_{jkS\alpha}$ associated to $H_0 : \Delta_j \mathcal{E}(S) \geqslant \Delta_k \mathcal{E}(S)$ to break ties. We adopt the test statistic approach since this seems more natural for breaking potential multi-way ties.

The tests will not be used in a conventional way (where $T_{jS\alpha} = 1$ indicates indicate that under repeated sampling under the null hypothesis, the current sample is unlikely.) Rather, they are simply a tool for determining the next action in a model selection procedure. The utility in taking this perspective is that many properties of hypothesis tests happen to be the same as those required for the general model selection procedure described below.

Throughout this discussion, we assume that such a feasible set of hypothesis tests exists and satisfies certain properties outlined below. We then provide an example giving primitive conditions on a linear model with heteroskedastic disturbances for which the general forward testing results apply.

We will then define a model selection procedure which yields a subset $\widehat{S} \subset \{1, ..., p\}$. Following model selection, we turn our attention to studying the properties of the *post-forward-selection-estimator* defined by

$$\widehat{\theta} \in \arg \min_{\theta : \text{supp}(\theta) \subset \widehat{S}} \ell_{f_\theta}(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n).$$

and our goal is to study the risk properties of this estimator. To summarize, the algorithm for forward selection given the set of hypothesis tests $\{T_{jS\alpha}, W_{jS}\}$ is given formally by:

Algorithm 1: Testing-Based Forward Selection

Initialize. Set $\widehat{S} = \{\}$.
For $1 \leqslant k \leqslant p$**:**
If: $T_{j\widehat{S}\alpha} = 1$ for some $j \in \{1, ..., p\} \setminus \widehat{S}$, then for

$$\widehat{j} \in \arg\max \left\{ W_{j\widehat{S}} : T_{j\widehat{S}\alpha} = 1 \right\},$$

Update: $\widehat{S} = \widehat{S} \cup \{\widehat{j}\}$.
Else: Break.
Set: $\widehat{\theta} \in \arg\min_{\theta:\mathrm{supp}(\theta)\subset\widehat{S}} \ell_{f_\theta}(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n)$
Set: $\widehat{f}(\cdot) = f_{\widehat{\theta}}(\cdot)$

3. Formal Conditions

This section formally states conditions on the hypothesis tests conditions on the data before analyzing properties of Algorithm 1. These conditions are measures of the quality of the given testing procedure and the regularity of the data. These measures defined in the below conditions are sufficient for proving useful performance bounds on the post-forward-selection estimator.

Condition 1 [*Data and Sparsity*]. Fix n. $(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ are distributed according to P. There is a set $S^* \subset \{1, ..., p\}$ with $|S^*| = s$ and a constant c_1 such that

$$\mathcal{E}(S^*) \leqslant c_1.$$

Condition 2 [*Hypothesis Tests*]. There are tests $T_{jS\alpha} \in \{0, 1\}$, test statistics W_{jS} determined by the data $(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n)$. There are constants c_2, c_2', c_2'' and for each $N \leqslant p$ there is $\delta_2 = \delta_2(N)$ such that each of the following conditions hold:

(I) The tests have power in the sense that with probability $1 - \delta_2$,

$$T_{jS\alpha} = 1 \ \text{ for every } j, |S| \leqslant N, \ \text{ such that } -\Delta_j \mathcal{E}(S) \geqslant c_2.$$

(II) The tests control size in the sense that probability of the event

$$T_{jS\alpha} = 1 \text{ for some } j, |S| \leqslant N \text{ such that } -\Delta_j \mathcal{E}(S) \leqslant c_2'$$

is no more than $\alpha + \delta_2$.

(III) With probability $1-\delta_2$,

$$W_{jS} \geqslant W_{kS} \text{ if and only if} - \Delta_j \mathcal{E}(S) \geqslant -c_2'' \Delta_k \mathcal{E}(S)$$

for each $j, k, |S| \leqslant N$, provided $T_{jS\alpha} = T_{kS\alpha} = 1$.

Condition 3 [*Sparse Eigenvalues*]. The components of $\psi_k(\cdot)$ are normalized so that

$$\frac{1}{n}\sum_{i=1}^{n} \psi_k^2(x_i) = 1$$

for every $1 \leqslant k \leqslant p$. Denote by $\psi_S(x_i)$ the vector with components $\psi_k(x_i)$, $k \in S$. For each $N \leqslant p$ there are constants $c_3 = c_3(N)$ and $\delta_3 = \delta_3(N)$ such that with probability $1-\delta_3(N)$,

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\psi_S(x_i)\psi_S(x_i)'\right)^{-1}, \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} \psi_S(x_i)\psi_S(x_i)'\right)^{-1} \leqslant c_3$$

for any S with $|S| \leqslant N$.

Condition 4 [*Estimation Quality*]. The infinum $\inf_{f\in\mathbf{F}} \mathrm{E}\ell_f$ is attained at f^* and the infinum $\inf_{\mathrm{supp}(\theta)\subset S^*} \mathrm{E}\ell_{f_\theta}$ is attained at θ^*. Define $\epsilon_i := y_i - f^*(x_i)$ and $a_i = f^*(x_i) - f_{\theta^*}(x_i)$. For $S \subset \{1, ..., p\}$, the infinum $\inf_{\mathrm{supp}(\theta)\subset S} \mathrm{E}\ell_{f_\theta}$ is attained at θ_S^* and $\epsilon_{iS} : y_i - f_{\theta_S^*}(x_i)$. The variables $\{y_i\}_{i=1}^n$ are normalized so that $\mathrm{E}\frac{1}{n}\sum_{i=1}^n y_i^2 = 1$. There is a constant c_4, for which with probability $1-\delta_4$ the following bounds all hold:

$$\max_{1\leqslant j\leqslant p} |\frac{1}{n}\sum_{i=1}^{n}\psi_j(x_i)\epsilon_i|, |\frac{1}{n}\sum_{i=1}^{n} f^*(x_i)\epsilon_i|, |\frac{1}{n}\sum_{i=1}^{n} f_{\theta^*}(x_i)\epsilon_i| \leqslant c_4$$

$$\max_{j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n} a_i\psi_j(x_i) - \mathrm{E}a_i\psi_j(x_i)\right| \leqslant c_4$$

$$\max_{j,l\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_j(x_i)\psi_l(x_i) - \mathrm{E}\psi_j(x_i)\psi_l(x_i)\right| \leqslant c_4.$$

In addition, for each $N \leqslant p$ there are constants $c_4' = c_4'(N)$ and $\delta_4' = \delta_4'(N)$ such that with probability at least $1-\delta_4'$, the following bounds hold:

$$\max_{S:|S|\leqslant N,\ \mathcal{E}(S)-\mathcal{E}(S^*)\leqslant 2sc_2c_3(N)} \ \max_{j\in S}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_j(x_i)(\epsilon_{iS} - \epsilon_i)\right| \leqslant c_4'.$$

Condition 1 is asserts that there is a sparse set $|S^*|$ with $\mathcal{E}(S^*)$ less than c_1. This set need not be unique. A common assumption in high dimensional modelling is the existence of a sparse set of useful predictors. This formulation measures simultaneously the number of covariates needed (s) to get within a target level (c_1) of population loss.

Condition 2 defines parameters that measure the quality of a given set of hypothesis tests. The constants measure quantities related to the size and power of the tests in the event they were used to assess statistical significance. These measure provide a convenient language in which to discuss the properties of the tests, but we emphasise here that the hypothesis tests considered *should not* be thought of as providing a measure of statistical significant. More accurately, they are simply a tool for model selection which coincidentally have many properties in common with traditional hypothesis tests. We hope that the common properties will allow the large past theory in hypothesis testing to find an alternative use in a large set of model selection problems.

Condition 3 is a sparse eigenvalue condition useful for proving results about high dimensional techniques like Lasso. In standard regression analysis where the number of covariates is small relative to the sample size, a conventional assumption used in establishing desirable properties of conventional estimators of θ is that $\frac{1}{n}\sum_{i=1}^n \psi(x_i)\psi(x_i)'$ has full rank. In the high dimensional setting, will be singular if $p > n$ and may have an ill-behaved inverse even when $p \leqslant n$. However, good performance of the Lasso estimator only requires good behavior of certain moduli of continuity of $\frac{1}{n}\sum_{i=1}^n \psi(x_i)\psi(x_i)'$. There are multiple formalizations and moduli of continuity that can be considered in establishing the good performance of Lasso; see [11]. We focus our analysis on a simple eigenvalue condition that is suitable for most econometric applications which was used in [5]. Condition SE could be shown to hold under more primitive conditions by adapting arguments found in [6] which build upon results in [45] and [36]; see also [35].

Finally, Condition 4 is needed to measure the quality of the post-model selection estimation step. The normalization $\mathrm{E}\frac{1}{n}\sum_{i=1}^n y_i^2 = 1$ is imposed for convenience; it also implicitly assumes second moments for the sum of the random variables $\{y_i\}_{i=1}^n$. The ϵ_i should be considered as idiosyncratic disturbances and the constant c_4 is used to bound empirical correlations with the covariates. c_4 should be considered as a constant measuring the extent to which a central-limit-like result holds. The constants c_4' measure a similar quantity as c_4 but uniformly over a much larger set of averages. This would in principal drive c_4' to be much larger than c_4, however, the constraint on $\mathcal{E}(S) - \mathcal{E}(S^*)$ ensures that the variances of the terms $\epsilon_i - \epsilon_{iS}$ are much smaller than the variances of ϵ_i.

With the conditions in place, we state the main theorem of the paper. The purpose of the theorem is to provide a tool to understand performance properties of forward selection in applications. In the theorem, n should be considered fixed. Given $c_1, c_2, c_2', c_2'', c_3, c_4, c_4', \delta_2, \delta_3, \delta_4, \delta_4', \alpha$, define:

$$\delta = 3\delta_2((\mathcal{C}_2+1)s) - \delta_3((\mathcal{C}_2+1)s) - \delta_4 - \delta_4'((\mathcal{C}_2+1)s)$$
$$\mathcal{C}_1 = c_1 + 2c_4 + sc_2c_3(s) + 2\widehat{s}c_3(\widehat{s})c_4(c_4 + c_4'(\widehat{s}))$$
$$+ 2s\max\{c_1,c_2\}c_3(s+\widehat{s})c_4 + [2s\max\{c_1,c_2\}c_3(s+\widehat{s})]^2c_4$$

$\mathcal{C}_2$ is defined by the largest value of

$$C(m) = \left(K_G^{\mathbb{R}}\right)^2 C_1^{-2}(1 + C_2^{1/2} + C_2)^2 c_3(m+s)$$

in the first set of contiguous integers in $[1,n]$ which satisfies $m \leqslant C(m)s$. $K_G^{\mathbb{R}} < 1.783$ is Grothendieck's constant, and

$$C_1 = \min\left\{ c_3(m+s)^{1/2}\frac{({c_2'}^{1/2} - c_1^{1/2})_+}{c_2^{1/2} + c_1^{1/2}}, {c_2''}^{1/2}c_3(m+s)\left[\frac{{c_2'}^{1/2} - 2c_1^{1/2}}{({c_2'}^{1/2} - c_1^{1/2})_+}\right]_+ \right\}$$

and $C_2 := c_3(m+s)^{-1}C_1$.

The constants defined above are referenced in the statement of the theorem. C_1 and C_2 are constants controlling the ratio $\Delta_j\mathcal{E}(S)/\Delta_k\mathcal{E}(S)$ when j is selected before k for $j \notin S^*, k \in S^*$ in the case of C_1 and for $j \in S^*, k \in S^*$ in the case of C_2. With probability at least δ, $\mathcal{C}_1$ and $\mathcal{C}_2$ control the estimation error and the number of covariates selected into the final model. This is formalized in the following theorem.

Theorem 1. *Fix n. Suppose that the assumptions on all data $(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n)$ listed above in Conditions 1,3,4 hold. Suppose that the assumptions in Condition 2 hold for a set of tests $T_{jS\alpha}, W_{jS}$. Then the bounds*

$$\frac{1}{n}\sum_{i=1}^n (f^*(x_i) - f_{\widehat{\theta}}(x_i))^2 \leqslant \mathcal{C}_1$$
$$\widehat{s} \leqslant (\mathcal{C}_2+1)s$$

hold with probability at least $1-\alpha-\delta$.

Proof. The proof of Theorem 1 is deserving of its own section (Section 7) and is presented after an example and some additional discussion of practical implementation. □

Comment 3.1. The tests are assumed to a notion of family-wise error rate. A similar result is expected to hold under an analogous false discovery proportion assumption since this should in principal preserve the statement $\widehat{s} \leqslant (\mathcal{C}_2+1)s$ up to a multiplicative constant.

Comment 3.2. The theorem provides a basis for understanding the prediction made by a model selected and fit by the Forward Selection Algorithm 1 described above. Below we give an example to a linear model with heteroskedastic data. We note that the theorem can be applied, at each n within a sequence $P^{(n)}$ of data generating processes. Under certain regularity conditions, we derive rates of type $O_{P^{(n)}}(s \log p/n)$ on the prediction norm and show that the constant in $\widehat{s} \leqslant (\mathcal{C}_2 + 1)s$ can be taken as to be $\mathcal{C}_2 = O(1)$. This gives convergence rates typical of those seen for Lasso and Post-Lasso.

4. Example: Heteroskedastic Disturbances

In this section we give an example of the use of Theorem 1 by illustrating an application of model selection in the presence of heteroskedasticity. We verify the primitive testing conditions set forth in Theorem 1 for a set of tests which are constructed based on the Heteroskedasticity-Consistent standard errors those described in [44]. We consider a sequence of data generating processes $P = P^{(n)}$. We will often omit dependence on n. We begin by outlining assumption on the data, and then provide exact details of the testing procedure. We focus on the linear model with fixed covariates.

Condition Ex1.1 [*Model*]. For each n the following model holds:

$$y_i = \psi(x_i)'\theta^* + \epsilon_i$$

with $x_i \in \mathcal{X} = \mathcal{X}_n$ deterministic and $\psi(\cdot) : \mathcal{X} \to \mathbb{R}^p$, with $p = p(n)$. Furthermore, ϵ_i are independent across i, not necessarily identically distributed, and have mean zero. Finally, $s = s(n) := |\text{supp}(\theta^*)|$.

The fact that the disturbances are not identically distributed and possibly heteroskedastic implies that classical iid standard errors may be inconsistent. Therefore, we adopt Huber-Eicker-White standard errors. In what follows, we describe in detail the testing procedure, before giving remaining formal regularity conditions, and finally proving a theorem about forward model selection in this setting.

Comment 4.1. Operating under the framework of fixed covariates is both convenient theoretically, and requires less stringent conditions on the data generating process. We give additional discussion of this issue after outlining the formal conditions.

We now describe the testing procedure. Still in the paradigm of quadratic loss, note that for any subset S and any $j \notin S$, the following two conditions are easily seen to be equivalent:

$$(1) \quad [\theta^*_{jS}]_j \neq 0 \qquad \text{and} \qquad (2) \quad \Delta_j \mathcal{E}(S) \neq 0$$

where θ^*_{jS} is defined the as optimal coefficient given the model $j \cup S$. We find it convenient to work with the formulation in condition (1). Consider the null hypothesis

$$H_0 : [\theta^*_{jS}]_j = 0$$

To test this hypothesis, construct test statistics based on the Heteroskedasticity-Consistent standard errors noted above. In order to do this, we construct the least squares estimate of θ^*_{jS}.

$$\widehat{\theta}_{jS} = \left[\frac{1}{n}\sum_{i=1}^{n}\psi_{jS}(x_i)\psi_{jS}(x_i)'\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\psi_{jS}(x_i)'y_i\right]$$

In addition, define

$$\widehat{\epsilon}_{ijS} = y_i - \psi_{jS}(x_i)'\widehat{\theta}_{jS}.$$

We next apply results on partial regression to construct our desired test. Let β_{jS} be the coefficient vector from the least squares regression of $\{\psi_j(x_i)\}_{i=1}^n$ on $\{\psi_k(x_i)\}_{i=1,k\in S}^n$. Consider the residuals from the previous regression, given by $\breve{\psi}_{jS}(x_i) = \psi_j(x_i) - \psi_S(x_i)'\beta_{jS}$. Then an estimate for $[\theta^*_{jS}]_j$ is given by

$$[\widehat{\theta}_{jS}]_j = \left[\frac{1}{n}\sum_{i=1}^{n}\breve{\psi}_{jS}(x_i)\breve{\psi}_{jS}(x_i)\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\breve{\psi}_{jS}(x_i)y_i\right].$$

and the heteroskedasiticty robust estimate of the variance

$$\widehat{V}_j = (\breve{\psi}'_{jS}\breve{\psi}_{jS})^{-1}\left[\sum_{i=1}^{n}\breve{\psi}_{jS}(x_i)^2\widehat{\epsilon}^2_{iS}\right](\breve{\psi}'_{jS}\breve{\psi}_{jS})^{-1}$$

Finally, define the test statistics:

$$W_{jS} = \widehat{V}_{jS}^{-1/2}\left|[\widehat{\theta}_{jS}]_j\right|.$$

We reject the null H_0 for large values of W_{jS} defined relative to an appropriately chosen threshold. To define the threshold first let $\eta_{jS} := (1\ ,\ \beta'_{jS})'$ be the coefficient vector for writing the residual $\breve{\psi}_j(x_i)$ in terms of $\psi_j(x_i), \psi_S(x_i)$. Without loss of generality, assume that the components of η_{jS} are nonnegative. Next, let $\Psi^{\widehat{\epsilon}}$ be defined so that $[\Psi^{\widehat{\epsilon}}_{jS}]_{k,l} = \sum_{i=1}^n \widehat{\epsilon}^2_{ijS}\psi_k(x_i)\psi_l(x_i)$ for $k, l \in jS$. Then define

$$\widehat{\tau}_{jS} = \frac{\eta_{jS}' \mathrm{diag}(\Psi_{jS}^{\widehat{\epsilon}})}{\sqrt{\eta_{jS}' \Psi_{jS}^{\widehat{\epsilon}} \eta_{jS}}}.$$

The term $\widehat{\tau}_{jS}$ will be helpful in addressing the fact that many different model selection paths are possible under different realizations of the data under P. Not taking this fact into account can potentially lead to false discoveries. We are in a position to state precisely the hypothesis tests $T_{jS\alpha}$.

Condition Ex1.2 [*Hypothesis Tests*]. Fix a tuning parameter $c_\tau > 1$ which is independent of n and a sequence of thresholds $\alpha = \alpha(n) \to 0$. The test statistics W_{jS} take the form described in the immediately preceding text. Furthermore, using the definition of $\widehat{\tau}_{jS}$ we assign:

$$T_{jS\alpha} = 1 \iff W_{js} \geqslant c_\tau \widehat{\tau}_{jS} \Phi^{-1}(1 - \alpha/p).$$

Comment 4.2. The $\Phi^{-1}(1 - \alpha/p)$ can be informally thought of as a Bonferonni correction term which takes into account of the fact that there are p potential covariates. The term $c_\tau \widehat{\tau}_{jS}$ can be informally thought of as a correction term which can account for the fact that the set S is random and can have many potential realizations. In the main simulations, we set $c_\tau = 1$ and we use $\alpha = .05$ for the sample sizes $n = 100$ and $n = 200$. We preliminary trials (not reported below) also tried $c_\tau = 1$ and noted that this choice does not seem to affect the quality of selected models. Below, we do report simulations which use other, less conservative thresholds for significance. With forward model selection, we find that using less conservative thresholds in fact slightly improves performance. However, all of the theoretical results presented in this paper address only the threshold stated above.

Having described the testing procedure, we record regularity conditions on the datagenerating process.

Condition Ex1.3 [*Sparse Eigenvalues and Irrepresentability*]. Let N_n be a sequence such that $N_n/s \to \infty$. For each S such that $|S| \leqslant N_n$,

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} \psi_S(x_i)\psi_S(x_i)'\right)^{-1} = O(1)$$

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 \psi_S(x_i)\psi_S(x_i)'\right)^{-1} = O(1) \text{ with probability } 1 - o(1).$$

In addition, for η_{jS} defined as above, let $c_{\mathrm{irr}} = \max_{j,|S| \leqslant N_n} \|\eta_{jS}\|_1$. Then $c_{\mathrm{irr}} = O(1)$.

Condition Ex1.4 [*Regularity*]. $(\psi(x_i)'\theta^*)^2 = O(1)$ uniformly for each $i = 1, ..., n$ and for each n. The disturbances ϵ_i satisfy

$$\max_{i \leqslant n} \mathrm{E}\epsilon_i^2 = O(1), \quad \max_{j \leqslant p} \frac{\left(\sum_{i=1}^n \mathrm{E}|\psi_j(x_i)^3 \epsilon_i^3|\right)^{1/3}}{\left(\sum_{i=1}^n \mathrm{E}\psi_j(x_i)^2 \epsilon_i^2\right)^{1/2}} = O(n^{-1/6})$$

For each subset $|S| \leqslant N_n$, let ϵ_{iS} be defined as earler. Decompose $\epsilon_{iS} = \epsilon_i + \xi_{iS}$. Then with probability $1 - o(1)$, the following large deviation result holds:

$$\left| \frac{1}{n} \sum_{i=1}^n \breve{\psi}_{jS}(x_i)^2 \epsilon_i \xi_{iS} \right| \leqslant \frac{1}{n} \sum_{i=1}^n \breve{\psi}_{jS}(x_i)^2 \xi_{iS}^2 \quad \text{for each } j \leqslant p, \ |S| \leqslant N_n.$$

Finally, we have the rate conditions:

$$\frac{N_n^2 \log^2 p}{n} \to 0, \quad \frac{\log^3 p}{n} \to 0.$$

Condition 1 describes the model and Condition 2 describes the testing procedure. The terms in the threshold are $\Phi^{-1}(1 - \alpha/p)$, which should be thought of as a Bonferroni multiple testing correction; and $c_\tau \widehat{\tau}_{j\widehat{S}}$ are needed as a correction for the fact that the sets $\widehat{S}$ are random.

Condition 3 gives conditions on the sparse eigenvalues of certain key matrices. Finally, Condition 3 assume an irrepresentability condition which may be strong in some cases. [39], [46] assume that $c_{\text{irr}} < 1$. In addition, [31] use an analogous assumption to $c_{\text{irr}} = O(1)$ in the context of learning high dimensional graphs. Below we make note of how to adjust the coming theorem when this condition is violated.

Condition 4 states regularity conditions on ϵ_i which are useful for proving central limit theorems and laws of large numbers. Note that ξ_{ijS} defined in Condition 4 are functions only of $\psi(x_i)$. The condition $\max_{j \leqslant p} \frac{\left(\sum_{i=1}^n \mathrm{E}|\psi_j(x_i)^3 \epsilon_i^3|\right)^{1/3}}{\left(\sum_{i=1}^n \mathrm{E}\psi_j(x_i)^2 \epsilon_i^2\right)^{1/2}} = O(n^{-1/6})$ allows the use of moderate deviation bounds for self-normalized sums. A more primitive condition is that $\mathrm{E}\psi_j(x_i)^2 \epsilon_i^2$ are bounded uniformly away from zero and above, and $\mathrm{E}\psi_j(x_i)^3 \epsilon_i^3$ are bounded uniformly above. We use the higher level condition since the covariates are fixed and we wish to allow $\psi_j(x_i)^2$ to be arbitrarily small or zero for some observations i. Finally, the two rate conditions provide bounds on the relative sizes of s, p, n since $s < N_n$.

Given the above assumptions, we have the following theorem which calculates convergence rates for the greedy testing-based forward selection procedure.

Theorem 2. *Under sequences $P = P^{(n)}$ and tests $T_{jS\alpha}, W_{jS}$ which satisfy the Conditions Ex1.1, Ex1.2, Ex1.3, Ex1.4, Algorithm 1 produces a model fit such that*

$$\frac{1}{n}\sum_{i=1}^{n}(f_{\theta^*}(x_i) - f_{\widehat{\theta}}(x_i))^2 = O_P(s \log p/n)$$

and $\widehat{s} = O(s)$ with probability $1 - o(1)$.

Proof. The proof of the Theorem 2 follows from a verification of the conditions of Theorem 1 and explicit calculation of the required constants. The details of the proof are given in the appendix.

□

Comment 4.3. We suspect that an analogous result holds for dependent data and HAC-type estimation (see [33], [2].) The required central limit results are beyond the scope of this work, though we mention that using the moderate deviation results of [17] we can already construct a feasible testing-based forward model selection procedure. Cluster-type standard errors for large-T-large-n and fixed-T-large-n panels can be used by adapting arguments from [7].

Comment 4.4. The procedure is conservative in that it applies a correction resembling a Bonferroni correction to maintain desired size properties. Given the current analysis, it is unclear theoretically whether lower thresholds (for example step-down thresholds) can be used. Simulation results resented in the next section suggest that step-down procedures actually perform better than the tests outlined above in most of the settings considered.

Comment 4.5. The condition $c_{\text{irr}} = O(1)$ is potentially restrictive (see discussion above). If instead the unrestrictive condition $c_{\text{irr}} = O(\sqrt{s})$ holds, then the following similar result can be shown: $\frac{1}{n}\sum_{i=1}^{n}(f_{\theta^*}(x_i) - f_{\widehat{\theta}}(x_i))^2 = O_P(s^2 \log p/n)$ and $\widehat{s} = O(1)s$.

5. Simulation

The results in the previous sections suggest that estimation with Forward Regression should produce good results in large sample sizes. In this section we simulate several different data generating processes to evaluate the performance of the Forward selection estimator. We compare the estimates to that of Lasso and Post-Lasso since these are popular and important generic high dimensional estimation strategies.

We consider the following data generating process:

$$y_i = x_i'\theta + \epsilon_i, \; i = 1, ..., n$$
$$p = \dim(x_i) = c_p n, \quad \theta_j = b^{j-1}$$
$$x_{ij} \sim N(0,1), \text{ with } \mathrm{corr}(x_{ij}, x_{ik}) = .5^{|j-k|}$$
$$\epsilon_i \sim \sigma_i N(0,1), \; \sigma_i = \exp(\rho \sum_{j=1}^{p} .75^{(p-j)} x_{ij}).$$

We replicate all simulations with parameter choices

$$b \in \{.75, .5, -.5, -.75\},$$
$$\rho \in \{0, .5\}$$
$$c_p \in \{.5, 2\}.$$

The parameter b controls the sparseness of the problem; for instance, when $b = .75$ the problem is more dense than when $b = .5$. The parameter ρ controls the amount of heteroskedasticity in the data, so that $\rho = 0$ means iid observations and $\rho = .5$ means heteroskedastic. Finally, we consider simulations where the number of explanatory variables is both less than the sample size ($c_p = .5$) and more than the sample size ($c_p = 2$).

In order to construct the test statistics, we use a both classical IID standard errors as well Huber-Eicker-White standard errors and compare the performance of the resulting estimators. We assess the size θ_j^* by comparing $[\widehat{\theta}_{jS}]_j/\text{s.e.}([\widehat{\theta}_{jS}]_j)$ to each of three thresholds τ_{jS}. First, we use the threshold described the paper given by $c_\tau \widehat{\tau}_{jS} \Phi^{-1}(1 - \alpha/p)$ with $c_\tau = 1$, $\alpha = .05$. The resulting estimator is called Forward I. Second, we use simply a Bonferroni correction $\Phi^{-1}(1 - .\alpha/p)$ with $\alpha = .05$. The resulting estimator is called Forward II. Finally, we use a step down threshold where, at any juncture with working model S, we use the threshold $\Phi^{-1}(1 - \alpha/(p - |S|))$. This estimator is called Forward III.

To construct a Lasso and Post-Lasso estimate, we use the implementation found in [5]. Their implementation chooses penalty loadings for each covariate based on an in sample measure of the variability of the covariate-specific score. They require two tuning parameters which are directly analogous to c_τ and α, so we again use $c_\tau = 1.1$ and $\alpha = .05$. Finally, we consider an infeasible estimator, which selects a model consisting of $\{j : |\theta_j^*| > 1/\sqrt{n}\}$.

The results are presented in Tables 1-8 in the appendix. Though neither Forward Selection, nor Lasso dominate the other in all simulations, there are important instances when the forward selection estimators consistently outperform the Lasso-based estimators. Forward selection estimates tend to do better relative to Post-Lasso in the presence of heteroskedasticity. The general pattern is that in the presence of heteroskedasticity, the use of Huber-Eicker-White standard errors substantially improves performance.

In addition, Lasso and Post-Lasso give very poor estimates when $b = -.5$ and $b = -.75$, while the forward selection estimators perform well (relative to Oracle). This suggests that the performance of these estimators depends on the configuration of the signal, not just the relative size of the signal to the noise. Finally, the Forward II and Forward III estimators seem to perform better than the Forward I estimator in general, suggesting that the proposed thresholds are possibly too conservative.

6. Empirical Illustration: Estimating the effects of Institutions on Economic Output

In order to illistrate the use of testing-based forward model selection to help answer an empirical economic question, we revisit the problem of estimating the effect of institution quality on aggregate economic output considered by Acemoglu, Johnson, and Robinson in [1]. A similar exercise on this data using Lasso-based methods was performed in [8].

To estimate the effect of institutions on output, it is necessary to address the fact that *both* (1) better institutions can lead to higher output; and (2) higher output can also lead to the development of better institutions. Because institutions and output levels both potentially affect each other, a simple correlation or regression analysis will not recover the causal quantity of interest. [1] introduce an instrumental variable strategy, using early European settler mortality as an instrument for institution quality. The validity of this instrument requires first a relevance assumption that early settler mortality is predictive of quality of current institutions. [1] argue that settlers set up lasting institutions in places where they were more likely to establish long term settlements. They cite several references documenting the fact that Europeans were acutely aware of mortality rates in their colonies. They also note that the institutions set up by early European settlers tend to be highly persistent. These arguments make the relevance assumption likely to hold. The exclusion restriction assumption is justified in [1] by the argument that GDP, while persistent, is unlikely to be strongly influenced by mortality rates centuries ago, except through institutions.

In their paper, [1] note that their IV strategy will be invalid if there are other factors that are highly persistent and related to the development of institutions within a country and to the countrys GDP. The primary candidate for such a factor discussed in [1] is geography. In this exercise, we take as given the fact that after controlling adequately for geography, it is possible to use their instrument strategy to correctly identify the effect of institutions on output. The outstanding problem then becomes the question of how, exactly, to adequately control for geography. [1] controlled for the distance from the equator in their baseline specification. They also considered specifications with continent dummies; see Table 4 in [1] .

In principal, there are many ways to construct control variables related to a broad notion such as geography. These may include variables based on temperature, yearly rain fall, or terrain. In this exercise, we construct a large set of different geographic variables. We then use testing based-forward model selection to choose from among the many variables and perform a subsequent IV analysis. Let x_i be a country level variable with components consisting of the dummy variables for Africa, Asia, North America, and South America plus the variables latitude, latitude^2, latitude^3, (latitude $-$ .08)$_+$, (latitude$-$.16)$_+$, (latitude$-$.24)+, ((latitude$-$.08)$_+$)^2, ((latitude$-$.16)$_+$)^2, ((latitude $-$ .24)$_+$)^2, ((latitude $-$ .08)$_+$)^3, ((latitude $-$ .16)$_+$)^3, ((latitude $-$ .24)$_+$)^3 where latitude denotes the distance of a country from the equator normalized to be between 0 and 1 which is the same set of controls as in [8]. Consider the model:

$$\log(\text{GDP per capita}_i) = \text{Protection from Expropriation}_i\theta + x_i'\beta + \epsilon_i$$

Here, "Protection from Expropriation" is the same as was used in [1]: a measure of the strength of individual property rights that is used as a proxy for the strength of institutions. We use the same set of 64 country-level observations as [1]. When the set of control variables for geography, x_i, is flexible enough, it is guaranteed that nothing can be learned about the effect of interest, θ, because of lack of statistical precision. [1] do not encounter such a problem because they assume the effect of geography is adequately captured by one variable. Using forward selection, we present a complimentary analysis which chooses controls from among our constructed set of geographic variables. We now describe the model selection procedure, which proceeds in several steps in order to ensure robustness against possible model selection mistakes.

Consider the fully expanded set of structural equations. This gives the following three relations:

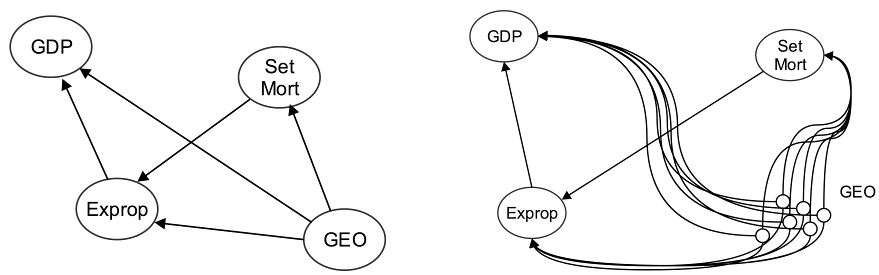
$$\begin{aligned} log(\text{GDP per capita}_i) &= \text{Protection from Expropriation}_i\theta + x_i'\beta + \epsilon_i \\ \text{Protection from Expropriation}_i &= \text{Settler Mortality}_i\pi_1 + x_i'\Pi_2 + v_i \\ \text{Settler Mortality}_i &= x_i'\gamma + u_i \end{aligned}$$

which yields three reduced form equations relating the structural variables to the controls:

$$\begin{aligned} log(\text{GDP per capita}_i) &= x_i'\beta + \tilde{\epsilon}_i \\ \text{Protection from Expropriation}_i &= x_i'\tilde{\Pi}_2 + \tilde{v}_i \\ \text{Settler Mortality}_i &= x_i'\gamma + u_i. \end{aligned}$$

The problem is represented pictorally in Figure 1. The left graph is a representation of the equations listed above. The right graph demonstrates that our desire to include a variable for geography can be done with many different "geography" control variables. The lack of an arrow between settler mortality and GDP highlights our exclusion restriction assumption.

Figure 1.



By arguments similar to those given in [8], in conjuction with the types of bounds reported in Section 5, it can be shown robust inference for θ after model selection over the variables constructed in x_i is possible. To accomplish this we can take the union of the set of variables selected by running testing-based forward selection on each of the three reduced form equations. We summarize this procedure below.

Algorithm 2: Estimating the effect of institution quality on aggregate economic output

Step 1. Use testing-based forward model selection over

$$log(\text{GDP per capita}_i) = x_i'\beta + \tilde{\epsilon}_i$$

Set: $\widehat{S}_1 = \{\text{Selected Covariates}\}$

Step 2. Use testing-based forward model selection over

$$\text{Protection from Expropriation}_i = x_i'\tilde{\Pi}_2 + \tilde{v}_i$$

Set: $\widehat{S}_2 = \{\text{Selected Covariates}\}$

Step 3. Use testing-based forward model selection over

$$\text{Settler Mortality}_i = x_i'\gamma + u_i$$

Set: $\widehat{S}_3 = \{\text{Selected Covariates}\}$

Step 4. Set: $\widehat{S} = \widehat{S}_1 \cup \widehat{S}_2 \cup \widehat{S}_3$

Run standard IV regression using $\widehat{S}$ as the set of controls.

Equivalently, we select all geographic variables which have a statistically non-negligible effect on any of the three variables: $log(\text{GDP per capita}_i)$, Protection from Expropriation$_i$, Settler Mortality$_i$. Valid estimation and inference of the structural parameter, θ, can then proceed by conventional IV estimation. Note importantly, that because three model selection steps will be used, the final estimates are robust to classical concerns about pre-test biases.

In Table 1 we present our estimates. The first column of the table labeled "Latitude" gives baseline results that control linearly for latitude which corresponds to the findings of [1] suggesting a strong positive effect of improved institutions on output with a reasonably strong first-stage. The second columns controls for all 16 of the constructed geography variables. This yields a visibly imprecise estimate of the effect of interest. This is expected, since the number of control variables, 16, is large enough relative to the sample size, 64, to prohibit precise estimation. The last column of Table 1 labeled "Forward Selection" controls for the union of the set of variables selected by running testing-based forward selection on each of the three reduced form equations, using heteroskedasticity-consistent standard errors and significance thresholds as described in Section 5. The last column is simply the IV estimate of the structural equation with the Africa dummy and the selected latitude spline term as the control variables. Interestingly, the results are qualitatively similar to the baseline results though the first-stage is somewhat weaker and the estimated structural effect is slightly smaller.

Table 1.

	Latitude	All Controls	Forward Selection
First Stage	-0.5372	-0.2182	-0.3802
	(0.1545)	(0.2011)	(0.1686)
Structural Estimate	0.9692	0.9891	0.8349
	(0.2128)	(0.8005)	(0.3351)

Selected variables: 1_{Africa}, $(\text{latitude} - .16)1_{\text{latitude}>.16}$

7. Proof of Theorem 1

Proof. The proof of this theorem has two main steps. First we bound the prediction norm on the event that the number of selected covariates, $\widehat{s}$ is less than N for N determined later. This part of the proof follows a similar outline to the proof of performance bounds of Post-Lasso, like those given in [5]. The second part of the proof requires a bound on the number of selected covariates $\widehat{s}$ and requires different theoretical methods than those used previously to analyse high dimensional problems; in particular, we must keep closer track of information on the relative magnitudes of all coefficients. We now begin the proof. In order to ease exposition, but still ensure completeness, we will defer routine calculations to a supplimentary appendix.

Let $\theta^*_{\widehat{S}} := \arg\min_{\mathrm{supp}(\theta)\subset \widehat{S}} \mathrm{E}\ell_{f_\theta}$. We introduce the notation $\ell(\theta) = \ell_{f_\theta}(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n)$. Also, define $\epsilon_i = y_i - f^*(x_i)$, $a_i := f^*(x_i) - x_i'\theta^*$. It will also in the course of the proof be convenient to define the following symbol for functions $\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ provided it exists: $\langle g, h\rangle = \mathrm{E}\frac{1}{n}\sum_{i=1}^n g(x_i,y_i)h(x_i,y_i)$. For vectors and matrices of functions we use the same symbol and apply it element-wise so that $\langle [g_{jk}],[h_{jk}]\rangle = [\langle g_{jk}, h_{jk}\rangle]$.

By definition of $\widehat{\theta}$, it follows that $\ell(\widehat{\theta}) \leqslant \ell(\theta^*_{\widehat{S}})$. Expanding the quadratics, $\ell(\widehat{\theta}), \ell(\theta^*_{\widehat{S}})$, and following Calculation 1 in the appendix, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (f_{\theta^*}(x_i) - f_{\widehat{\theta}}(x_i))^2 &\leqslant |\mathcal{E}(\widehat{S}) - \mathcal{E}(S^*)| + |2\frac{1}{n}\sum_{i=1}^n \epsilon_i\psi(x_i)'(\widehat{\theta} - \theta^*_{\widehat{S}})| \\ &+ \left|\frac{1}{n}\sum_{i=1}^n (a_i\psi(x_i) - \mathrm{E}a_i\psi(x_i))'(\theta^*_{\widehat{S}} - \theta^*)\right| \\ &+ \left|\frac{1}{n}\sum_{i=1}^n (\psi(x_i)'(\theta^*_{\widehat{S}} - \theta^*))^2 - \mathrm{E}(\psi(x_i)'(\theta^*_{\widehat{S}} - \theta^*))^2\right| \\ &:= D_1 + D_2 + D_3 + D_4\end{aligned}$$

The terms on the right hand side are bound separately. With probability $1-\delta_2(N)$, (for N sufficiently large and chosen later), Algorithm 1 terminates at a step with

$$-\Delta_j\mathcal{E}(\widehat{S}) \leqslant c_2$$

for every $j \notin \widehat{S}$. Because of the structure of quadratic loss, the quantity $\Delta_j\mathcal{E}(\widehat{S})$ is directly related to the change in R^2 (defined conventionally). This allows an application of the results of [18], Lemma 3.3, which relate the increase in R^2 from inclusion of a set of regressors to the increase in R^2 from inclusion of each regressor from the set separately. Noting that $|S^*\setminus\widehat{S}| \leqslant s$ and applying [18] yields

$$|\mathcal{E}(\widehat{S}^*) - \mathcal{E}(\widehat{S})| \leqslant c_3(s+\widehat{s})\sum_{j\in S^*\setminus\widehat{S}} -\Delta_j\mathcal{E}(S) \leqslant sc_2c_3(s+\widehat{s}).$$

Next to construct a bound for D_2, note that by Hölder's inequality,

$$\left|\frac{1}{n}\sum_{i=1}^{n}2\epsilon_i\psi(x_i)'(\widehat{\theta}-\theta^*_{\widehat{S}})\right| \leqslant \|\frac{1}{n}\sum_{i=1}^{n}2\epsilon_i\psi(x_i)\|_\infty\|\widehat{\theta}-\theta^*_{\widehat{S}}\|_1$$

Use Condition 4 to bound $\|\frac{1}{n}\sum_{i=1}^n 2\epsilon_i\psi(x_i)\|_\infty \leqslant c_4$ with probability $1-\delta_4$. For any subset $S \subset \{1,...,p\}$, let ψ_S be the matrix in $\mathbb{R}^{n\times|S|}$ with elements $\psi_j(x_i)$ for $j \in S$. Using the calculation in the appendix, the following bounds hold:

$$\|\widehat{\theta}-\theta^*_{\widehat{S}}\|_1 \quad \leqslant \widehat{s}c_3(\widehat{s})(c_4+c_4'(N))$$

Finally, similar bounds can be constructed for D_3 and D_4 and, as detailed in the appendix, we have

$$|D_3| \leqslant 2s\max\{c_1,c_2\}c_3(s+\widehat{s})c_4$$
$$|D_4| \leqslant [2s\max\{c_1,c_2\}c_3(s+\widehat{s})]^2c_4.$$

The fact that $\frac{1}{n}\sum_{i=1}^n(f_{\theta^*}(x_i)-f^*(x_i))^2 \leqslant c_1+2c_4$ with probability $1-\delta_4$, together with $c_1+2c_4+D_1+D_2+D_3+D_4 \leqslant \mathcal{C}_1$ and taking $N=(\mathcal{C}_2+1)s$ yield that with probability at least $1-\alpha-\delta$,

$$1_{\{\widehat{s}\leqslant(\mathcal{C}_2+1)s\}}\cdot\frac{1}{n}\sum_{i=1}^{n}(f^*(x_i)-f_{\widehat{\theta}}(x_i))^2 \leqslant \mathcal{C}_1.$$

We next prove the probabilistic bound for the size of the selected set $\widehat{s}$ in terms of s. In the course of this proof, it eases exposition to talk about "true and false regressors" so we introduce a few conventions and notations. Let v_k, $k=1,...,s$ denote "true regressors" which are defined as random variables realized as vectors in $\mathbb{R}^n$ with components $\{\psi_k(x_i)\}_{i=1}^n$ with $k \in S^*$, ordered according to the order they are selected into the model (any unselected regressors can be ordered arbitrarily and placed at the end of the list). Let $\tilde{v}_1,...,\tilde{v}_s$ be orthogonalized regressors obtained from $v_1,...,v_s$ through the Gram-Schmidt process, with respect to $\langle\cdot,\cdot\rangle$ define above. We use the normalization that $\langle\tilde{v}_k,\tilde{v}_k\rangle=1$.

We define "false regressors" simply as those which do not belong to S^*. Suppose there are m "falsely chosen" regressors $w_1,...,w_m$, ie. regressors chosen from the complement of S^*. Let $\tilde{w}_j$ denote orthogonalized versions of w_j (we define the corresponding normalization later), where the orthogonalization order is defined with respect to the previously selected regressors, including the true regressors.

Let $\tilde{V}=[\tilde{v}_1,...,\tilde{v}_s]$. Note then that there is $\tilde{\theta}\in\mathbb{R}^s$ such that $\tilde{V}\tilde{\theta} = [v_1,...,v_s]\theta^*$. In addition, each $\tilde{w}_j$ can be decomposed into components $\tilde{w}_j = \tilde{r}_j+\tilde{u}_j$ with $\tilde{r}_j \in \text{span}(\tilde{V})$ and $\tilde{u}_j \in \text{span}(\tilde{V})^\perp$. Importantly, we assume that $\tilde{w}_j$ is normalized so that $\langle\tilde{u}_j,\tilde{u}_j\rangle=1$. Furthermore, $\tilde{r}_j$ can be expressed as a linear combination $\tilde{V}\tilde{\gamma}_j$ with $\gamma\in\mathbb{R}^s$, and we will often simply identify $\tilde{\gamma}_j$

with w_j. Finally, let $a_i := f^*(x_i) - f_{\theta^*}(x_i)$ and a the vector with components a_i. A simple derivation (see the calculation in the appendix) can be made to show that the incremental decrease in empirical loss from the jth false selection is

$$-\Delta_j \mathcal{E}(S_{j-1}) = \frac{(\tilde{\gamma}_j' \tilde{\theta} + \langle \tilde{w}_j, a \rangle)^2}{\langle \tilde{r}_j, \tilde{r}_j \rangle}$$

Therefore, the quantity $\tilde{\gamma}_j' \tilde{\theta}$ is closely related to the jth false selection.

The key point which we argue next is that if there are C_1 and C_2 such that

$$\tilde{\gamma}_j' \tilde{\theta} / \tilde{\theta}_k \geqslant C_1 > 0 \text{ and } \tilde{\theta}_k / \tilde{\theta}_l \geqslant C_2 > 0$$

for all $j, k, l > k$ then a bound can be given on the number of false selections in terms of C_1, C_2. We prove this fact first, then later derive values for C_1 and C_2 which hold with high probability.

The idea guiding the following argument is that if too many variables are selected, then they must be correlated with each other. Informally, this is motivated by transitivity, since by merit of being selected, they must be correlated to $f^*(x_i)$. For a discussion of partial transitivity of correlation, see [37]. This transitivity, once made formal, together with the sparse eigenvalue assumption will lead to a contradiction. To make this logic precise, let $\tilde{W} = [\tilde{w}_1, ..., \tilde{w}_m]$, and similarly decompose $\tilde{W} = \tilde{R} + \tilde{U}$. Then $\langle \tilde{W}, \tilde{W} \rangle = \langle \tilde{R}, \tilde{R} \rangle + \langle \tilde{U}, \tilde{U} \rangle$. Since $diag(\langle \tilde{U}' \tilde{U} \rangle) = I$, it follows that the average correlation between the $\tilde{u}_j$, given by $\bar{\rho} := \frac{1}{m(m-1)} \sum_{j \neq l} \langle \tilde{u}_j, \tilde{u}_l \rangle$, must be bounded below by

$$\bar{\rho} \geqslant -\frac{1}{m-1}$$

due to the positive definiteness of $\langle \tilde{U}, \tilde{U} \rangle$. This implies an upper bound on the average off-diagonal term in $\langle \tilde{R}, \tilde{R} \rangle$ since $\langle \tilde{W}, \tilde{W} \rangle$ is a diagonal matrix. More explicitly, since $\tilde{v}_k$ are orthonormal, we have that the sum of all the elements of $\langle \tilde{R}, \tilde{R} \rangle$ is given by $\| \sum_{j=1}^m \tilde{\gamma}_j \|_2^2$. Since $\| \sum_{j=1}^m \tilde{\gamma}_j \|_2^2 = \sum_{j=1}^m \| \tilde{\gamma}_j' \|_2^2 + \sum_{j \neq l} \tilde{\gamma}_j' \tilde{\gamma}_l$ and since $\langle \tilde{W}, \tilde{W} \rangle$ is a diagonal matrix, it must be the case that $\sum_{j \neq l} \tilde{\gamma}_j' \tilde{\gamma}_l = \bar{\rho}$. Therefore,

$$\bar{\rho} = \frac{1}{m(m-1)} \left(\Big\| \sum_{j=1}^m \tilde{\gamma}_j \Big\|_2^2 - \sum_{j=1}^m \| \tilde{\gamma}_j \|_2^2 \right) \leqslant \frac{1}{m-1}$$

Note that $\|\tilde{\gamma}_j\|^2 \leqslant c_3(m+s) - 1$ since by Condition 3, $\langle \tilde{w}_j, \tilde{w}_j \rangle / \langle \tilde{u}_j, \tilde{u}_j \rangle \leqslant c_3(m+s)$. This then implies that

$$\Big\| \sum_{j=1}^m \tilde{\gamma}_j \Big\|_2^2 \leqslant m c_3 (m+s)$$

We next calculate the constant C so that $\left\| \sum_{j=1}^{m} \tilde{\gamma}_j \right\|_2^2 \geqslant mc_3(m+s)$ whenever $m \geqslant Cs$. Intuitively, the idea is to apply a bound like the Cauchy-Schwarz inequality in reverse to obtain

$$\left\| \sum_{j=1}^{m} \tilde{\gamma}_j \right\|_2^2 \|\tilde{\theta}\|_2^2 \geqslant \sum_{j=1}^{m} \tilde{\gamma}_j' \tilde{\theta}$$

and use what we know about $\tilde{\gamma}_j' \tilde{\theta}$ (given selection for w_j into the model) to derive a *lower* bound for $\| \sum \tilde{\gamma}_j \|_2^2$.

This bound is useful for illustrating the main idea, however, it is not tight enough for the present purpose, unless a very restrictive β-min condition is imposed. Instead, the argument relies on Grothendieck's inequality which is a theorem of functional analysis proven by Alexander Grothendieck in 1953 ([20], see for a review, [10]) which bounds the $\|\Gamma\|_{\infty \to 1}$ of the matrix Γ (defined below) which can then be related to $\| \sum_{j=1}^{m} \tilde{\gamma}_j \|_2^2$.

We define the following matrices. Let $m_1, ..., m_s$ be sets with m_k containing those j such that w_j is selected before v_k, but not before any other true regressor. Let

$$\Gamma = \begin{pmatrix} \sum_{j \in m_1} \tilde{\gamma}_{js} & \sum_{j \in m_1} \tilde{\gamma}_{j2} & \cdots & \sum_{j \in m_1} \tilde{\gamma}_{js} \\ 0 & \sum_{m_2} \tilde{\gamma}_{j2} & \cdots & \sum_{j \in m_2} \tilde{\gamma}_{js} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{j \in m_s} \tilde{\gamma}_{js} \end{pmatrix}$$

note that the kth row of Γ is equal to $\sum_{j \in m_k} \tilde{\gamma}_j$ since the orthogonalization process had enforced $\tilde{\gamma}_{jl} = 0$ for each $l < k$. Next let

$$B = \begin{pmatrix} \frac{\tilde{\theta}_1}{\tilde{\theta}_1} & \frac{\tilde{\theta}_2}{\tilde{\theta}_1} & \cdots & \frac{\tilde{\theta}_s}{\tilde{\theta}_1} \\ \frac{\tilde{\theta}_2}{\tilde{\theta}_1} & \frac{\tilde{\theta}_2}{\tilde{\theta}_2} & \cdots & \frac{\tilde{\theta}_s}{\tilde{\theta}_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\tilde{\theta}_s}{\tilde{\theta}_1} & \frac{\tilde{\theta}_s}{\tilde{\theta}_2} & \cdots & \frac{\tilde{\theta}_s}{\tilde{\theta}_s} \end{pmatrix}$$

Observe that the diagonal elements of the product satisfy the equality

$$[\Gamma B]_{k,k} = \sum_{j \in m_k} \tilde{\gamma}_j \tilde{\theta}/\tilde{\theta}_k.$$

by the condition of false selection, this implies that

$$[\Gamma B]_{k,k} \geqslant C_1.$$

Further observe that whenever $\tilde{\theta}_k \geqslant C_2 \tilde{\theta}_l$ for each $k, l > k$, assuming without loss of generality that $C_2 \leqslant 1$, we have $B + C_2^{-1} I \in \mathcal{M}_G^+ := \{Z \in \mathbb{R}^{s \times s} : Z \geqslant 0, \operatorname{diag}(Z) \leqslant 1\}$. This can be checked by constructing auxiliary random variables who have covariance matrix $B + C_2^{-1} I$: inductively build a covariance matrix where the $(k+1)$th random variable has θ_k covariance with the kth random variable. Then $B + C_2^{-1} I$ has a positive definite symmetric matrix square root so hat $D^2 = B + C_2^{-1} I$. Therefore, $B = (D + C_2^{-1/2} I)(D - C_2^{-1/2} I)$. Note that the rows (and columns) of D each have norm $\leqslant 1 + C_2^{-1}$ and therefore B decomposes into a product $B = E'F$ where the rows of E, F all have norm bounded by $1 + C_2^{-1} + C_2^{-1/2} =: C_2'$.

Consider the set

$$\mathcal{M}_G = \{Z \in \mathbb{R}^{s \times s} : Z_{ij} = X_i' Y_j \text{ for some } X_i, Y_j \in \mathbb{R}^s, \|X_i\|_2, \|Y_j\|_2 \leqslant 1\}$$

and observe that

$$\bar{B} := {C_2'}^{-1} B \in \mathcal{M}_G.$$

Then this observation allows the use of Grothendieck's inequality (for which we use the exact form described in [21]) which gives

$$\operatorname{tr}(\Gamma \bar{B}) \leqslant \max_{Z \in \mathcal{M}_G} \operatorname{tr}(\Gamma Z) \leqslant K_G^{\mathbb{R}} \|\Gamma'\|_{\infty \to 1}.$$

Here, $K_G^{\mathbb{R}}$ is an absolute constant called Grothendieck's constant. It is known to be less than 1.783. Therefore, we have $C_1 m \leqslant \operatorname{tr}(\Gamma B) = C_2' \operatorname{tr}(\Gamma \bar{B})$, which implies

$$\left(K_G^{\mathbb{R}}\right)^{-1} {C_2'}^{-1} C_1 m \leqslant \|\Gamma\|_{\infty \to 1}.$$

Therefore, there is $\nu \in \{-1, 1\}^s$ such that $\|\nu' \Gamma\|_1 \geqslant \left(K_G^{\mathbb{R}}\right)^{-1} {C_2'}^{-1} C_1 m$. For this particular choice of ν, it follows that

$$\|\nu' \Gamma\|_2 \geqslant s^{-1/2} \left(K_G^{\mathbb{R}}\right)^{-1} {C_2'}^{-1} C_1 m$$

Without loss of generality (due to the ambiguity of assigning signs to $\tilde{w}_j$ in the orthogonalization process), we may assume that $\nu_j = 1$ for each $j \leqslant s$. Then

$$\|1' \Gamma\|_2^2 = \| \sum_j \tilde{\gamma}_j \|_2^2$$

Since from before, we had noted that $\left\| \sum_{j=1}^{m} \tilde{\gamma}_j \right\|_2^2 \leqslant mc_3(m+s)$, it follows that

$$s^{-1} \left(K_G^{\mathbb{R}} \right)^{-2} {C_2'}^{-2} C_1^2 m^2 \leqslant mc_3(m+s)$$

which yields the conclusion

$$m \leqslant c_3(m+s) C_1^{-2} {C_2'}^{2} \left(K_G^{\mathbb{R}} \right)^2 s.$$

This proves that if $\tilde{\gamma}_j' \tilde{\theta} / \tilde{\theta}_k \geqslant C_1$ and $\tilde{\theta}_k / \tilde{\theta}_l \geqslant C_2$ for all $k, l > k$ then we have a bound on the number of falsely chosen regressors in terms of C_1 and C_2.

In the appendix we show that the constants given in the statement of Theorem 1 are sufficient. This concludes the proof of Theorem 1.

□

8. Conclusion

This paper developes theory for testing-based forward model selection in linear regression problems. We prove bounds on the performance of greedy stepwise regression which include probabilistic bound on prediction error and number of selected covariates. We verify that the stated regularity conditions on the set of hypothesis tests are attained for the linear model under fixed covariates and heteroskedastic disturbances. We compare the performance of Lasso and Post-Lasso to the performance of Forward Selection in Simulation studies and find that in many instances, Forward Selection shows better performance.

References

[1] Daron Acemoglu, Simon Johnson, and James A. Robinson. The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5):1369–1401, 2001.

[2] Donald W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, 1991.

[3] J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146:304–317, 2008.

[4] J. Bai and S. Ng. Boosting diffusion indices. *Journal of Applied Econometrics*, 24, 2009.

[5] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80:2369–2429, 2012. Arxiv, 2010.

[6] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. ArXiv, 2009.

[7] A. Belloni, V. Chernozhukov, C. Hansen, and D. Kozbur. Inference in high dimensional panel models with an application to gun contorl. *ArXiv:1411.6507*, 2014.

[8] A. Belloni, C. Hansen, and V. Chernozhukov. High-dimensional mehods and inference on structural and treatmentt effectcs. *Journal of Economic Perspectives*, 28:29–50, 2014.

[9] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection amongst high-dimensional controls with an application to abortion on crime. *Review of Economic Studies*, 81(2):608–650, 2014.

[10] C. Alan Bester and Christian Hansen. Grouped effects estimators in fixed effects models. *Journal of Econometrics*, 2014. forthcoming.

[11] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[12] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[13] F. Bunea, A. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[14] F. Bunea, A. B. Tsybakov, , and M. H. Wegkamp. Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006) (G. Lugosi and H. U. Simon, eds.)*, pages 379–391, 2006.

[15] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[16] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[17] X. Chen, Q.-M. Shao, and W. Biao Wu. Self-normalized Cram\'er Type Moderate Deviations under Dependence. *ArXiv e-prints*, September 2014.

[18] Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1057–1064, New York, NY, USA, 2011. ACM.

[19] Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[20] Alexander Grothendieck. Resume de la theorie metrique des produits tensoriels topologiques. *Soc. Mat. Sao-Paulo*, 8:1 – 79, 1953.

[21] O. Guédon and R. Vershynin. Community detection in sparse networks via Grothendieck's inequality. *ArXiv e-prints*, November 2014.
[22] Christian Hansen and Damian Kozbur. Instrumental variables estimation with many weak instruments using regularized {JIVE}. *Journal of Econometrics*, 182(2):290 – 308, 2014.
[23] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2009.
[24] Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 2010.
[25] Bing-Yi Jing, Qi-Man Shao, and Qiying Wang. Self-normalized cramr-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215, 2003.
[26] Keith Knight. Shrinkage estimation for nearly singular designs. *Econometric Theory*, 24:323–337, 2008.
[27] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincar Probab. Statist.*, 45(1):7–57, 2009.
[28] D. Kozbur. Inference in Additively Separable Models With a High Dimensional Set of Conditioning Variables. *ArXiv e-prints*, March 2015.
[29] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.
[30] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *arXiv:0903.1468v1 [stat.ML]*, 2010.
[31] N. Meinshausen and P. Buhlmann. High dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34:1436–1462, 2006.
[32] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):2246–2270, 2009.
[33] Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.
[34] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *arXiv:0812.2818v1 [math.ST]*, 2008.
[35] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *ArXiv:1106.1151*, 2011.
[36] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:10251045, 2008.
[37] T. Tao. When is correlation transitive?, June 2014.
[38] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288, 1996.
[39] Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10):2231–2242, Oct 2004.
[40] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.
[41] Sara van de Geer, Peter Bhlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 06 2014.
[42] M. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, May 2009.
[43] Hansheng Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104:488:1512–1524, 2009.
[44] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[45] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.
[46] Tong Zhang. On the consistency of feature selection using greedy least squares. *Journal of Machine Learning*, 10:555–568, 2009.
[47] Tong Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Trans. Inf. Theor.*, 57(7):4689–4708, July 2011.

Appendix A. Tables

Table 1. Forward Model Selection Simulation Results:

Sample Size : $n = 100$, Dimensionality : $p = .5n$
Disturbances : Homoskedastic, Replications : 1000

	MPEN	MSSS	MPEN	MSSS
	Classical S.E.		White S.E.	
	A. $\theta_j = .75^{j-1}$			
Forward I	0.85	2.95	0.86	2.99
Forward II	0.55	4.87	0.55	4.90
Forward III	0.54	4.92	0.55	4.94
Lasso	3.10	4.05	2.78	4.25
Post-Lasso	0.66	4.05	0.63	4.25
Oracle	0.33	9.00	0.33	9.00
	A. $\theta_j = .5^{j-1}$			
Forward I	0.45	1.79	0.46	1.80
Forward II	0.36	2.28	0.36	2.38
Forward III	0.36	2.29	0.36	2.39
Lasso	1.17	1.35	1.40	1.75
Post-Lasso	0.57	1.35	0.44	1.75
Oracle	0.21	4.00	0.21	4.00
	C. $\theta_j = (-.5)^{j-1}$			
Forward I	0.41	1.04	0.41	1.06
Forward II	0.33	1.57	0.33	1.64
Forward III	0.33	1.58	0.33	1.64
Lasso	0.89	0.00	0.83	0.19
Post-Lasso	0.89	0.00	0.79	0.19
Oracle	0.19	4.00	0.19	4.00
	D. $\theta_j = (-.75)^{j-1}$			
Forward I	0.69	1.16	0.69	1.17
Forward II	0.54	2.25	0.54	2.29
Forward III	0.54	2.27	0.54	2.31
Lasso	1.02	0.01	0.99	0.10
Post-Lasso	1.01	0.01	0.98	0.10
Oracle	0.30	9.00	0.30	9.00

Note: We print mean prediction error norm (MPEN) and mean size of selected set (MSSS) for several estimators. Forward I implements the testing based forward selection procedure in the text with $c_\tau = 1.1$, $\alpha = .05$ for both iid and White standard errors. Forward II alters thresholds to simply $\Phi^{-1}(1-\alpha/p)$, resembling a Bonferroni correction. Forward III alters the threshold of $\Phi^{-1}(1-\alpha/(p-|\widehat{S}|))$, resembling a step-down procedure. Lasso and Post-Lasso use the implementation in [5] with tuning parameters $c_\tau = 1.1, \alpha = .05$. Oracle is infeasible estimator and selects the model $\{j : |\theta_j^*| > 1/\sqrt{n}\}$.

Table 2. Forward Model Selection Simulation Results:

Sample Size : $n = 100$, Dimensionality : $p = .5n$
Disturbances : Heteroskedastic, Replications : 1000

	MPEN	MSSS	MPEN	MSSS
	Classical S.E.		White S.E.	
	A. $\theta_j = .75^{j-1}$			
Forward I	1.59	1.22	1.47	1.34
Forward II	1.52	1.56	1.34	1.79
Forward III	1.52	1.56	1.34	1.80
Lasso	3.43	10.58	3.10	10.81
Post-Lasso	1.76	10.58	1.80	10.81
Oracle	1.04	9.00	1.03	9.00
	A. $\theta_j = .5^{j-1}$			
Forward I	1.06	0.82	0.95	0.83
Forward II	1.06	0.89	0.93	0.93
Forward III	1.06	0.90	0.93	0.93
Lasso	2.51	8.32	2.29	8.62
Post-Lasso	1.65	8.32	1.72	8.62
Oracle	0.66	4.00	0.68	4.00
	C. $\theta_j = (-.5)^{j-1}$			
Forward I	0.91	0.33	0.77	0.33
Forward II	0.92	0.35	0.78	0.35
Forward III	0.92	0.35	0.78	0.35
Lasso	1.98	7.25	1.77	7.12
Post-Lasso	1.73	7.25	1.69	7.12
Oracle	0.69	4.00	0.65	4.00
	D. $\theta_j = (-.75)^{j-1}$			
Forward I	1.07	0.31	0.97	0.25
Forward II	1.07	0.33	0.97	0.28
Forward III	1.07	0.33	0.97	0.28
Lasso	2.01	6.81	1.92	7.83
Post-Lasso	1.77	6.81	1.89	7.83
Oracle	1.03	9.00	1.06	9.00

Note: We print mean prediction error norm (MPEN) and mean size of selected set (MSSS) for several estimators. Forward I implements the testing based forward selection procedure in the text with $c_\tau = 1.1$, $\alpha = .05$ for both iid and White standard errors. Forward II alters thresholds to simply $\Phi^{-1}(1 - \alpha/p)$, resembling a Bonferroni correction. Forward III alters the threshold of $\Phi^{-1}(1 - \alpha/(p - |\widehat{S}|))$, resembling a step-down procedure. Lasso and Post-Lasso use the implementation in [5] with tuning parameters $c_\tau = 1.1, \alpha = .05$. Oracle is infeasible estimator and selects the model $\{j : |\theta_j^*| > 1/\sqrt{n}\}$.

Table 3. Forward Model Selection Simulation Results:

Sample Size : $n = 100$, Dimensionality : $p = 2n$
Disturbances : Homoskedastic, Replications : 1000

	MPEN	MSSS	MPEN	MSSS
	Classical S.E.		White S.E.	
	A. $\theta_j = .75^{j-1}$			
Forward I	0.96	2.51	0.94	2.61
Forward II	0.62	4.34	0.61	4.45
Forward III	0.61	4.35	0.61	4.46
Lasso	2.55	3.63	2.43	4.00
Post-Lasso	0.74	3.63	0.67	4.00
Oracle	0.33	9.00	0.33	9.00
	A. $\theta_j = .5^{j-1}$			
Forward I	0.52	1.56	0.52	1.57
Forward II	0.38	2.14	0.39	2.25
Forward III	0.38	2.15	0.39	2.25
Lasso	1.07	1.12	1.13	1.58
Post-Lasso	0.68	1.12	0.49	1.58
Oracle	0.21	4.00	0.21	4.00
	C. $\theta_j = (-.5)^{j-1}$			
Forward I	0.41	1.03	0.41	1.06
Forward II	0.35	1.41	0.36	1.50
Forward III	0.36	1.41	0.36	1.50
Lasso	0.90	0.00	0.87	0.07
Post-Lasso	0.90	0.00	0.85	0.07
Oracle	0.20	4.00	0.19	4.00
	D. $\theta_j = (-.75)^{j-1}$			
Forward I	0.72	1.02	0.72	1.06
Forward II	0.59	1.83	0.60	1.91
Forward III	0.59	1.83	0.60	1.91
Lasso	1.02	0.00	1.01	0.06
Post-Lasso	1.02	0.00	1.00	0.06
Oracle	0.30	9.00	0.30	9.00

Note: We print mean prediction error norm (MPEN) and mean size of selected set (MSSS) for several estimators. Forward I implements the testing based forward selection procedure in the text with $c_\tau = 1.1$, $\alpha = .05$ for both iid and White standard errors. Forward II alters thresholds to simply $\Phi^{-1}(1 - \alpha/p)$, resembling a Bonferroni correction. Forward III alters the threshold of $\Phi^{-1}(1 - \alpha/(p - |\widehat{S}|))$, resembling a step-down procedure. Lasso and Post-Lasso use the implementation in [5] with tuning parameters $c_\tau = 1.1, \alpha = .05$. Oracle is infeasible estimator and selects the model $\{j : |\theta_j^*| > 1/\sqrt{n}\}$.

TABLE 4. Forward Model Selection Simulation Results:

Sample Size : $n = 100$, Dimensionality : $p = 2n$
Disturbances : Heteroskedastic, Replications : 1000

	MPEN	MSSS	MPEN	MSSS
	Classical S.E.		White S.E.	
	A. $\theta_j = .75^{j-1}$			
Forward I	1.65	1.01	1.55	1.12
Forward II	1.59	1.27	1.46	1.42
Forward III	1.59	1.28	1.46	1.42
Lasso	3.72	17.41	3.47	19.34
Post-Lasso	2.41	17.41	2.62	19.34
Oracle	1.04	9.00	1.06	9.00
	A. $\theta_j = .5^{j-1}$			
Forward I	1.12	0.70	1.00	0.74
Forward II	1.12	0.73	0.99	0.80
Forward III	1.12	0.73	0.99	0.80
Lasso	2.93	14.92	2.74	15.30
Post-Lasso	2.33	14.92	2.40	15.30
Oracle	0.67	4.00	0.68	4.00
	C. $\theta_j = (-.5)^{j-1}$			
Forward I	0.92	0.19	0.81	0.22
Forward II	0.93	0.20	0.82	0.23
Forward III	0.93	0.20	0.82	0.23
Lasso	2.39	12.66	2.34	13.77
Post-Lasso	2.30	12.66	2.38	13.77
Oracle	0.66	4.00	0.66	4.00
	D. $\theta_j = (-.75)^{j-1}$			
Forward I	1.10	0.17	0.99	0.16
Forward II	1.10	0.18	0.99	0.16
Forward III	1.10	0.18	0.99	0.16
Lasso	2.69	14.52	2.41	14.13
Post-Lasso	2.58	14.52	2.52	14.13
Oracle	1.08	9.00	1.04	9.00

Note: We print mean prediction error norm (MPEN) and mean size of selected set (MSSS) for several estimators. Forward I implements the testing based forward selection procedure in the text with $c_\tau = 1.1$, $\alpha = .05$ for both iid and White standard errors. Forward II alters thresholds to simply $\Phi^{-1}(1 - \alpha/p)$, resembling a Bonferroni correction. Forward III alters the threshold of $\Phi^{-1}(1 - \alpha/(p - |\widehat{S}|))$, resembling a step-down procedure. Lasso and Post-Lasso use the implementation in [5] with tuning parameters $c_\tau = 1.1, \alpha = .05$. Oracle is infeasible estimator and selects the model $\{j : |\theta_j^*| > 1/\sqrt{n}\}$.

Table 5. Forward Model Selection Simulation Results:

Sample Size : $n = 200$, Dimensionality : $p = .5n$
Disturbances : Homoskedastic, Replications : 1000

	MPEN	MSSS	MPEN	MSSS
	Classical S.E.		White S.E.	
	A. $\theta_j = .75^{j-1}$			
Forward I	0.64	4.03	0.64	4.04
Forward II	0.41	5.98	0.41	5.98
Forward III	0.41	6.00	0.41	6.00
Lasso	4.61	4.11	4.29	4.39
Post-Lasso	0.65	4.11	0.60	4.39
Oracle	0.25	10.00	0.25	10.00
	A. $\theta_j = .5^{j-1}$			
Forward I	0.35	2.12	0.36	2.11
Forward II	0.26	2.81	0.26	2.84
Forward III	0.26	2.82	0.26	2.84
Lasso	1.34	1.31	2.04	1.75
Post-Lasso	0.58	1.31	0.43	1.75
Oracle	0.16	4.00	0.16	4.00
	C. $\theta_j = (-.5)^{j-1}$			
Forward I	0.38	1.14	0.38	1.16
Forward II	0.24	2.04	0.24	2.07
Forward III	0.24	2.04	0.24	2.08
Lasso	0.89	0.00	0.87	0.06
Post-Lasso	0.89	0.00	0.86	0.06
Oracle	0.14	4.00	0.14	4.00
	D. $\theta_j = (-.75)^{j-1}$			
Forward I	0.63	1.47	0.63	1.48
Forward II	0.40	3.34	0.40	3.38
Forward III	0.40	3.35	0.40	3.39
Lasso	1.02	0.00	1.01	0.04
Post-Lasso	1.02	0.00	1.01	0.04
Oracle	0.22	10.00	0.22	10.00

Note: We print mean prediction error norm (MPEN) and mean size of selected set (MSSS) for several estimators. Forward I implements the testing based forward selection procedure in the text with $c_\tau = 1.1$, $\alpha = .05$ for both iid and White standard errors. Forward II alters thresholds to simply $\Phi^{-1}(1 - \alpha/p)$, resembling a Bonferroni correction. Forward III alters the threshold of $\Phi^{-1}(1 - \alpha/(p - |\widehat{S}|))$, resembling a step-down procedure. Lasso and Post-Lasso use the implementation in [5] with tuning parameters $c_\tau = 1.1, \alpha = .05$. Oracle is infeasible estimator and selects the model $\{j : |\theta_j^*| > 1/\sqrt{n}\}$.

Table 6. Forward Model Selection Simulation Results:

Sample Size : $n = 200$, Dimensionality : $p = .5n$
Disturbances : Heteroskedastic, Replications : 1000

	MPEN	MSSS	MPEN	MSSS
	Classical S.E.		White S.E.	
	A. $\theta_j = .75^{j-1}$			
Forward I	1.41	1.51	1.34	1.60
Forward II	1.27	2.10	1.19	2.19
Forward III	1.27	2.11	1.19	2.19
Lasso	4.37	11.65	3.95	12.77
Post-Lasso	1.46	11.65	1.58	12.77
Oracle	0.80	10.00	0.81	10.00
	A. $\theta_j = .5^{j-1}$			
Forward I	0.87	1.01	0.83	0.99
Forward II	0.85	1.15	0.78	1.15
Forward III	0.85	1.15	0.78	1.15
Lasso	2.89	8.68	2.79	9.97
Post-Lasso	1.33	8.68	1.44	9.97
Oracle	0.49	4.00	0.50	4.00
	C. $\theta_j = (-.5)^{j-1}$			
Forward I	0.79	0.47	0.68	0.50
Forward II	0.79	0.49	0.68	0.52
Forward III	0.79	0.49	0.68	0.52
Lasso	1.72	7.04	1.70	8.01
Post-Lasso	1.34	7.04	1.45	8.01
Oracle	0.49	4.00	0.49	4.00
	D. $\theta_j = (-.75)^{j-1}$			
Forward I	1.00	0.37	0.92	0.39
Forward II	1.00	0.42	0.92	0.44
Forward III	1.00	0.42	0.92	0.44
Lasso	1.89	7.67	1.78	8.16
Post-Lasso	1.55	7.67	1.58	8.16
Oracle	0.79	10.00	0.78	10.00

Note: We print mean prediction error norm (MPEN) and mean size of selected set (MSSS) for several estimators. Forward I implements the testing based forward selection procedure in the text with $c_\tau = 1.1$, $\alpha = .05$ for both iid and White standard errors. Forward II alters thresholds to simply $\Phi^{-1}(1 - \alpha/p)$, resembling a Bonferroni correction. Forward III alters the threshold of $\Phi^{-1}(1 - \alpha/(p - |\widehat{S}|))$, resembling a step-down procedure. Lasso and Post-Lasso use the implementation in [5] with tuning parameters $c_\tau = 1.1, \alpha = .05$. Oracle is infeasible estimator and selects the model $\{j : |\theta^*_j| > 1/\sqrt{n}\}$.

TABLE 7. Forward Model Selection Simulation Results:

Sample Size : $n = 200$, Dimensionality : $p = 2n$
Disturbances : Homoskedastic, Replications : 1000

	MPEN	MSSS	MPEN	MSSS
	Classical S.E.		White S.E.	
	A. $\theta_j = .75^{j-1}$			
Forward I	0.70	3.69	0.71	3.65
Forward II	0.44	5.59	0.45	5.56
Forward III	0.44	5.59	0.45	5.57
Lasso	3.89	3.77	3.80	4.05
Post-Lasso	0.71	3.77	0.66	4.05
Oracle	0.25	10.00	0.25	10.00
	A. $\theta_j = .5^{j-1}$			
Forward I	0.37	2.02	0.37	2.01
Forward II	0.29	2.58	0.29	2.62
Forward III	0.29	2.58	0.29	2.62
Lasso	1.07	1.07	1.56	1.59
Post-Lasso	0.69	1.07	0.47	1.59
Oracle	0.16	4.00	0.16	4.00
	C. $\theta_j = (-.5)^{j-1}$			
Forward I	0.40	1.06	0.40	1.08
Forward II	0.26	1.86	0.26	1.91
Forward III	0.26	1.86	0.26	1.91
Lasso	0.89	0.00	0.89	0.02
Post-Lasso	0.89	0.00	0.89	0.02
Oracle	0.14	4.00	0.14	4.00
	D. $\theta_j = (-.75)^{j-1}$			
Forward I	0.67	1.24	0.67	1.24
Forward II	0.44	2.96	0.44	2.96
Forward III	0.44	2.96	0.44	2.97
Lasso	1.02	0.00	1.02	0.01
Post-Lasso	1.02	0.00	1.02	0.01
Oracle	0.22	10.00	0.23	10.00

Note: We print mean prediction error norm (MPEN) and mean size of selected set (MSSS) for several estimators. Forward I implements the testing based forward selection procedure in the text with $c_\tau = 1.1$, $\alpha = .05$ for both iid and White standard errors. Forward II alters thresholds to simply $\Phi^{-1}(1 - \alpha/p)$, resembling a Bonferroni correction. Forward III alters the threshold of $\Phi^{-1}(1 - \alpha/(p - |\widehat{S}|))$, resembling a step-down procedure. Lasso and Post-Lasso use the implementation in [5] with tuning parameters $c_\tau = 1.1, \alpha = .05$. Oracle is infeasible estimator and selects the model $\{j : |\theta_j^*| > 1/\sqrt{n}\}$.

Table 8. Forward Model Selection Simulation Results:

Sample Size : $n = 200$, Dimensionality : $p = 2n$
Disturbances : Heteroskedastic, Replications : 1000

	MPEN	MSSS	MPEN	MSSS
	Classical S.E.		White S.E.	
	A. $\theta_j = .75^{j-1}$			
Forward I	1.47	1.29	1.41	1.40
Forward II	1.33	1.76	1.26	1.90
Forward III	1.33	1.76	1.26	1.90
Lasso	4.37	18.07	4.14	21.89
Post-Lasso	1.93	18.07	2.21	21.89
Oracle	0.79	10.00	0.83	10.00
	A. $\theta_j = .5^{j-1}$			
Forward I	0.89	0.92	0.85	0.92
Forward II	0.88	0.99	0.83	1.01
Forward III	0.88	0.99	0.83	1.01
Lasso	3.01	14.04	2.95	17.73
Post-Lasso	1.76	14.04	2.02	17.73
Oracle	0.49	4.00	0.49	4.00
	C. $\theta_j = (-.5)^{j-1}$			
Forward I	0.81	0.33	0.72	0.40
Forward II	0.81	0.34	0.72	0.42
Forward III	0.81	0.34	0.72	0.42
Lasso	2.00	11.97	2.11	14.80
Post-Lasso	1.76	11.97	1.96	14.80
Oracle	0.48	4.00	0.49	4.00
	D. $\theta_j = (-.75)^{j-1}$			
Forward I	1.01	0.22	0.95	0.27
Forward II	1.01	0.23	0.95	0.30
Forward III	1.01	0.23	0.95	0.30
Lasso	2.21	13.33	2.34	16.90
Post-Lasso	2.00	13.33	2.20	16.90
Oracle	0.79	10.00	0.81	10.00

Note: We print mean prediction error norm (MPEN) and mean size of selected set (MSSS) for several estimators. Forward I implements the testing based forward selection procedure in the text with $c_\tau = 1.1$, $\alpha = .05$ for both iid and White standard errors. Forward II alters thresholds to simply $\Phi^{-1}(1 - \alpha/p)$, resembling a Bonferroni correction. Forward III alters the threshold of $\Phi^{-1}(1 - \alpha/(p - |\widehat{S}|))$, resembling a step-down procedure. Lasso and Post-Lasso use the implementation in [5] with tuning parameters $c_\tau = 1.1, \alpha = .05$. Oracle is infeasible estimator and selects the model $\{j : |\theta_j^*| > 1/\sqrt{n}\}$.

Appendix B. Calculations

This appendix includes supporting calculations for the proof of the main result.

B.1. **Calculation 1.**

$$\sum_{i=1}^{n}(y_i - \psi(x_i)'\widehat{\theta})^2 \leqslant \sum_{i=1}^{n}(y_i - \psi(x_i)'\theta^*_{\widehat{S}})^2$$

$$\implies \sum_{i=1}^{n}(\psi(x_i)'\theta^* + \epsilon_i + a_i - \psi(x_i)'\widehat{\theta})^2 \leqslant \sum_{i=1}^{n}(\psi(x_i)'\theta^* + \epsilon_i + a_i - \psi(x_i)'\theta^*_{\widehat{S}})^2$$

$$\implies \sum_{i=1}^{n}(\psi(x_i)'\theta^* + \epsilon_i + a_i - \psi(x_i)'\widehat{\theta})^2 \leqslant \sum_{i=1}^{n}(\psi(x_i)'\theta^* + \epsilon_i + a_i - \psi(x_i)'\theta^*_{\widehat{S}})^2$$

$$\implies \sum_{i=1}^{n}[\psi(x_i)'(\theta^* - \widehat{\theta})]^2 + (\epsilon_i + a_i)^2 + 2(a_i + \epsilon_i)\psi(x_i)'(\theta^* - \widehat{\theta})$$

$$\leqslant \sum_{i=1}^{n}[\psi(x_i)(\theta^* - \theta^*_{\widehat{S}})]^2 + (\epsilon_i + a_i)^2 + 2(a_i + \epsilon_i)\psi(x_i)'(\theta^* - \theta^*_{\widehat{S}})$$

$$\implies \sum_{i=1}^{n}(f_{\theta^*}(x_i) - f_{\widehat{\theta}}(x_i))^2 \leqslant \sum_{i=1}^{n}[\psi(x_i)(\theta^* - \theta^*_{\widehat{S}})]^2 + 2(a_i + \epsilon_i)\psi(x_i)'(\widehat{\theta} - \theta^*_{\widehat{S}})$$

Considering $\widehat{S}$ fixed note that

$$\mathcal{E}(S) - \mathcal{E}(\widehat{S}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[(y_i - \psi(x_i)'\theta^*)^2 - \mathrm{E}(y_i - \psi(x_i)'\theta^*_{\widehat{S}})^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[(a_i + \epsilon_i)^2 - (a_i + \epsilon_i - \psi(x_i)'(\theta^*_{\widehat{S}} - \theta^*))^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[(\psi(x_i)'(\theta^*_{\widehat{S}} - \theta^*))^2 + 2(a_i + \epsilon_i)\psi(x_i)'(\theta^*_{\widehat{S}} - \theta^*)]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[(\psi(x_i)'(\theta^*_{\widehat{S}} - \theta^*))^2 + 2a_i\psi(x_i)'(\theta^*_{\widehat{S}} - \theta^*)]$$

$$\implies \sum_{i=1}^{n}(f_{\theta^*}(x_i) - f_{\widehat{\theta}}(x_i))^2 \leqslant \sum_{i=1}^{n}[\psi(x_i)'(\theta^* - \theta^*_{\widehat{S}})]^2 + 2(a_i + \epsilon_i)\psi(x_i)'(\widehat{\theta} - \theta^*_{\widehat{S}})$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}(f_{\theta^*}(x_i) - f_{\widehat{\theta}}(x_i))^2 \leqslant |\mathcal{E}(\widehat{S}) - \mathcal{E}(S^*)| + |2\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^*\psi(x_i)'(\widehat{\theta} - \theta^*_{\widehat{S}})|$$

$$+\left|\frac{1}{n}\sum_{i=1}^{n}(a_i\psi(x_i)-\mathrm{E}a_i\psi(x_i))'(\theta^*_{\widehat{S}}-\theta^*)\right|$$
$$+\left|\frac{1}{n}\sum_{i=1}^{n}(\psi(x_i)'(\theta^*_{\widehat{S}}-\theta^*))^2-\mathrm{E}(\psi(x_i)'(\theta^*_{\widehat{S}}-\theta^*))^2\right|$$
$$:= D_1+D_2+D_3+D_4$$

B.2. **Calculation 2.** Suppose that $|\mathcal{E}(\widehat{S})-\mathcal{E}(S^*)|\leqslant sc_2c_3(s)$ then we can bound D_3+D_4 by noting that

$$|\mathcal{E}(\widehat{S})-\mathcal{E}(S^*)|\equiv D_1\leqslant sc_2c_3(s)$$

implies a bound on $\|\theta^*_{\widehat{S}}-\theta^*\|_1$. To show this, define $d_{\widehat{S}}=\theta^*_{\widehat{S}}-\theta^*$. Recall that

$$\mathcal{E}(\widehat{S})-\mathcal{E}(S^*)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[(\psi(x_i)'(\theta^*_{\widehat{S}}-\theta^*))^2+2a_i\psi(x_i)'(\theta^*_{\widehat{S}}-\theta^*)]$$
$$=d'_{\widehat{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi(x_i)\psi(x_i)'\right]d_{\widehat{S}}+2\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}a_i\psi(x_i)'d_{\widehat{S}}$$

Consider two cases. First, if

$$\left|2\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}a_i\psi(x_i)'d_{\widehat{S}}\right|\leqslant\frac{1}{2}d'_{\widehat{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi(x_i)\psi(x_i)'\right]d_{\widehat{S}}$$

Then since the right hand side above is nonnegative, it follows that

$$D_1=\mathcal{E}(\widehat{S})-\mathcal{E}(S^*)\geqslant\frac{1}{2}d'_{\widehat{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi(x_i)\psi(x_i)'\right]d_{\widehat{S}}$$
$$\geqslant\frac{1}{2}\lambda_{\min}(\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi(x_i)\psi(x_i)')\|d_{\widehat{S}}\|_2^2$$

which implies that

$$\|d_{\widehat{S}}\|_1\leqslant|\sqrt{|\widehat{S}\cup S^*|}\frac{1}{\sqrt{2}}D_1^{1/2}\lambda_{\min}(\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi(x_i)\psi(x_i)')^{-1/2}$$

Consider the other case, that

$$|2\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}a_i\psi(x_i)'d_{\widehat{S}}|>\frac{1}{2}d'_{\widehat{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi(x_i)\psi(x_i)\right]d_{\widehat{S}}$$

Then bound

$$\left|2\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}a_i\psi(x_i)'d_{\widehat{S}}\right|\leqslant2\|d_{\widehat{S}}\|_1\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}a_i^2}\max_j\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi_j(x_i)^2}$$

Combining the above two bound with

$$\frac{1}{2}d'_{\widehat{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi(x_i)\psi(x_i)'\right]d_{\widehat{S}} \geqslant \lambda_{\min}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi(x_i)\psi(x_i)'\right]\|d_{\widehat{S}}\|_2^2$$

gives

$$\lambda_{\min}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi(x_i)\psi(x_i)'\right]\|d_{\widehat{S}}\|_2^2 \leqslant 2\|d_{\widehat{S}}\|_1\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}a_i^2}\max_j\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi_j(x_i)^2}$$

Simplifying by noting the assumed facts that $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\psi_j(x_i)2} = 1$ and $\lambda_{\min}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}x_{iS^*\cup\widehat{S}}x'_{iS^*\cup\widehat{S}}\right] \geqslant c_3(\widehat{s}+s)^{-1}$ yields

$$c_3(\widehat{s}+s)^{-1}(\widehat{s}+s)^{-1}\|d_{\widehat{S}}\|_1^2 \leqslant 2\|d_{\widehat{S}}\|_1 c_1$$

which implies that

$$\|d_{\widehat{S}}\|_1 \leqslant 2c_1c_3(s+\widehat{s})(\widehat{s}+s).$$

Summarizing the above calculation, we have that

$$\begin{aligned}\|d_{\widehat{S}}\|_1 &\leqslant \max\left\{2c_1c_3(s+\widehat{s})(s+\widehat{s}), \sqrt{\widehat{s}+s}2\sqrt{s}c_2c_3(\widehat{s}+s)^{1/2}c_3(\widehat{s}+s)^{1/2}\right\}\\ &= 2s\max\{c_1,c_2\}c_3(s+\widehat{s})\end{aligned}$$

Note now that

$$\begin{aligned}D_3 &\leqslant \|\theta^* - \theta^*_{\widehat{S}}\|_1\max_j\left|\frac{1}{n}\sum_{i=1}^{n}a_i\psi_j(x_i) - \mathrm{E}a_i\psi_j(x_i)\right|\\ &\leqslant 2s\max\{c_1,c_2\}c_3(s+\widehat{s})c_4, \quad \text{with probability} \quad 1-\delta(N)\\ D_4 &= \sum_{j,l}[\theta^* - \theta^*_{\widehat{S}}]_j[\theta^* - \theta^*_{\widehat{S}}]_l\frac{1}{n}\sum_{i=1}^{n}\psi_j(x_i)\psi_l(x_i)' - \mathrm{E}\psi_j(x_i)\psi_l(x_i)'\\ &\leqslant \|\theta^* - \theta^*_{\widehat{S}}\|_1^2\max_{j,l}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_j(x_i)\psi_l(x_i)' - \mathrm{E}\psi_j(x_i)\psi_l(x_i)'\right|\\ &\leqslant [2s\max\{c_1,c_2\}c_3(s+\widehat{s})]^2c_4, \quad \text{with probability} \quad 1-\delta(N)\end{aligned}$$

B.3. **Calculation 3.** In this subsection, we bound $2\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\psi(x_i)'(\widehat{\theta}-\theta^*_{\widehat{S}})$. Note that by Hölder's inequality,

$$\left|\frac{1}{n}\sum_{i=1}^{n}2\epsilon_i\psi(x_i)'(\widehat{\theta}-\theta^*_{\widehat{S}})\right| \leqslant \|\frac{1}{n}\sum_{i=1}^{n}2\epsilon_i\psi(x_i)\|_\infty\|\widehat{\theta}-\theta^*_{\widehat{S}}\|_1$$

Use Condition 4 to bound $\|\frac{1}{n}\sum_{i=1}^{n}2\epsilon_i\psi(x_i)\|_\infty \leqslant c_4$. Use the notation $\psi_{\widehat{S}}$ to be the matrix with elements $\psi_j(x_i)$ for $j \in \widehat{S}$, and similar for ψ_j. Define for each S, $\epsilon_{iS} = y_i - \psi_S(x_i)\theta^*_S$. Using Conditions 4 and 5, the following bounds hold:

$$\begin{aligned}
\|\widehat{\theta}-\theta^*_{\widehat{S}}\|_1 &= \|(\psi_{\widehat{S}}'\psi_{\widehat{S}})^{-1}\psi_{\widehat{S}}\epsilon_{\widehat{S}}\|_1 \\
&\leqslant \widehat{s}^{1/2}\left\|(\psi_{\widehat{S}}'\psi_{\widehat{S}})^{-1}\psi_{\widehat{S}}\epsilon_{\widehat{S}}\right\|_2 \\
&\leqslant \widehat{s}^{1/2}\lambda_{\min}(\frac{1}{n}\psi_{\widehat{S}}'\psi_{\widehat{S}})^{-1}\|\frac{1}{n}\psi_{\widehat{S}}'\epsilon_{\widehat{S}}\|_2 \\
&\leqslant \widehat{s}^{1/2}\lambda_{\min}(\frac{1}{n}\psi_{\widehat{S}}'\psi_{\widehat{S}})^{-1}\widehat{s}^{1/2}\max_{j\in\widehat{S}}|\frac{1}{n}\psi_j'\epsilon_{\widehat{S}}| \\
&\leqslant \widehat{s}^{1/2}\lambda_{\min}(\frac{1}{n}\psi_{\widehat{S}}'\psi_{\widehat{S}})^{-1}\widehat{s}^{1/2}\left(\max_{j\leqslant p}\left|\frac{1}{n}\psi_j'\epsilon\right| + \max_{j\in\widehat{S}}\left|\frac{1}{n}\psi_j'(\epsilon_{\widehat{S}}-\epsilon)\right|\right) \\
&\leqslant \widehat{s}c_3(\widehat{s})\left(c_4 + \max_{j\in\widehat{S}}\left|\frac{1}{n}\psi_j'(\epsilon_{\widehat{S}}-\epsilon)\right|\right) \\
&\leqslant \widehat{s}c_3(\widehat{s})\left(c_4 + \max_{S:|S|\leqslant N,\ \mathcal{E}(S)-\mathcal{E}(S^*)\leqslant 2sc_2c_3(N)}\ \max_{j\in S}\left|\frac{1}{n}\psi_j'(\epsilon_S-\epsilon)\right|\right) \\
&\leqslant \widehat{s}c_3(\widehat{s})(c_4+c_4'(N))
\end{aligned}$$

B.4. **Calculation 4.** Here we calculate $\Delta_j\mathcal{E}(S)$ in terms of the quantities $\tilde{\theta}, \tilde{\gamma}_j$ as defined in the main text. This is a simple exercise in applying first order conditions.

B.5. **Calculation 5.** Here we calculate the constants C_1 and C_2 in the proof of Theorem 1. To ease notation, we omit the dependence on N of the constant c_3. Any appearence of c_3 is meant as $c_3(s+m)$.

In order for a false selection of w_j to occur while v_k is the next unselected true regressor, it is necessarily the case that for the current standing selected set S,

$$T_{jS} = 1 \text{ and } W_{jS} \geqslant W_{kS} \text{ if } T_{kS} = 1$$

In the case that $T_{kS} = 0$, then because of Condition 3, projection of v_k to the space spanned by the previously selected regressors has length at least

c_3^{-1} in the direction $\tilde{v}_k$, which yields

$$c_3^{-1}\tilde{\theta}_k + \langle \tilde{v}_k, a\rangle < c_2^{1/2}.$$

Then with Cauchy-Schwarz, $|\langle \tilde{v}_k, a\rangle| \leqslant (\langle \tilde{v}_k, \tilde{v}_k\rangle)^{1/2}(c_1)^{1/2} = (c_1)^{1/2}$ gives

$$\tilde{\theta}_k \leqslant \frac{c_2^{1/2} + (c_1)^{1/2}}{c_3^{-1}}.$$

At the same time, since w_j was selected,

$$\frac{(\tilde{\gamma}_j'\tilde{\theta} + \langle \tilde{w}_j, a\rangle)^2}{\langle \tilde{w}_j, \tilde{w}_j\rangle} > c_2'.$$

This, along with the fact that $\frac{1}{n}\tilde{w}_j'\tilde{w}_j \leqslant c_3$ by consequence of Condition 3, gives

$$\tilde{\gamma}_j'\tilde{\theta} \geqslant \left[c_3^{1/2}{c_2'}^{1/2} - c_1^{1/2}c_3^{1/2}\right]_+.$$

This implies the relation

$$\tilde{\gamma}_j'\tilde{\theta}/\tilde{\theta}_k \geqslant c_3^{1/2}\frac{\left[{c_2'}^{1/2} - c_1^{1/2}\right]_+}{{c_2}^{1/2} + c_1^{1/2}} =: C_1'.$$

In the other case, where $T_{kS} = 1$, then

$$\frac{(\tilde{\gamma}_j'\tilde{\theta} + \langle \tilde{w}_j, a\rangle)^2}{\langle \tilde{r}_j, \tilde{r}_j\rangle} \geqslant c_2''(c_3^{-1}\tilde{\theta}_k + \langle \tilde{v}_k, a\rangle)^2$$

then

$$\tilde{\gamma}_j'\tilde{\theta}/\tilde{\theta}_k \geqslant {c_2''}^{1/2}\left[c_3^{-1} + \frac{1}{\tilde{\theta}_k}\langle \tilde{v}_k' a\rangle\right]_+$$

but since $T_{kS} = 1$ then $\tilde{\theta} \geqslant c_3^{-1}(c_2' - \langle \tilde{v}_k, a\rangle)$ and $\frac{1}{\tilde{\theta}_k} \leqslant \frac{c_3}{c_2' - \langle \tilde{v}_k' a\rangle} \leqslant \frac{c_3}{c_2' - c_1}$ which implies that

$$\gamma_j'\tilde{\theta}/\tilde{\theta}_k \geqslant {c_2''}^{1/2}\left[c_3 - \frac{c_3 c_1^{1/2}}{({c_2'}^{1/2} - c_1^{1/2})_+}\right]_+ =: C_1''$$

defining $C_1 = \min\{C_1', C_1''\}$, we have that

$$\gamma_j'\tilde{\theta}/\tilde{\theta}_k \geqslant C_1.$$

Finally, by similar logic as above, we may take $C_2 = C_1 c_3^{-1}$.

B.6. **Here we prove Theorem 2.** Use the stacking notation defined previously. Let $\mathcal{P}_S = \psi_S(\psi_S'\psi_S)^{-1}\psi_S'$ $\mathcal{M}_S = I - \mathcal{P}_S$. Then

$$[\widehat{\theta}_{jS}]_j = [\theta^*_{jS}]_j + (\psi_j'\mathcal{M}_S\psi_j)^{-1}\psi_j'\mathcal{M}_S\epsilon_{jS}.$$

Use $\breve{\psi}_{jS}$ to denote $\mathcal{M}_S\psi_j$. Under quadratic loss we have

$$\Delta_j\mathcal{E}(S) = \mathrm{E}\frac{1}{n}\sum_{i=1}^{n}\left[(y_i - x_i'\theta^*_S)^2 - (y_i - x_i'\theta^*_{jS})^2\right]$$

and a simple derivation gives (see notes from Schennach, ECON 301: Empirical Analysis I, The University of Chicago, Autumn 2009, page 38)

$$\begin{aligned}\Delta_j\mathcal{E}(S) &= [\theta^*_{Sj}]_j^2\left(\left[\left(\left[\frac{1}{n}\sum_{i=1}^{n}\psi(x_i)\psi(x_i)'\right]_{Sj,Sj}\right)^{-1}\right]_{jj}\right)^{-1}\\ &:= [\theta^*_{Sj}]_j^2 A_{jS}\end{aligned}$$

Let:

$$\zeta_{jS} := \breve{\psi}_{jS}\epsilon_{jS}$$

$$\Sigma_{jS} = \sum_{i=1}^{n}\breve{\psi}^2_{ijS}\epsilon_{ijS}{}^2, \quad \widehat{\Sigma}_{jS} = \sum_{i=1}^{n}\breve{\psi}^2_{ijS}\widehat{\epsilon}^2_{ijS}$$

$$V_{jS} := A_{js}^{-2}\Sigma_{jS}, \quad \widehat{V}_{jS} := A_{js}^{-2}\widehat{\Sigma}_{jS}$$

finally, we define the quantity $\xi_{ijS} : \epsilon_{ijS} = \epsilon_i + \xi_{ijS}$.

We denote by ϵ the vector of true disturbances (without subscripts). We use similar notation for ξ_{jS} etc. Then we can write

$$\widehat{[\theta_{jS}]}_j - [\theta^*_{jS}]_j = A_{jS}^{-1}\zeta_{jS}$$

We analyze the quantity

$$\begin{aligned}\widehat{V}_{jS}^{-1/2}(\widehat{[\theta_{jS}]}_j - [\theta_{jS}]^*_j) &= \widehat{V}_{jS}^{-1/2}A_{jS}^{-1}\zeta_{jS} = \widehat{\Sigma}_{jS}^{-1/2}\zeta_{jS}\\ &= \Sigma_{jS}^{-1/2}\zeta_{jS} + (\widehat{\Sigma}_{jS}^{-1/2} - \Sigma_{jS}^{-1/2})\zeta_{jS}\end{aligned}$$

The goal is to get control on the two terms on the right hand side uniformly for all $j, |S| \leqslant N_n$, for the sequence N_n defined in the Conditions above. Analyze the two terms on the right hand side above separately. Starting with the second:

$$\max_{jS} |(\widehat{\Sigma}_{jS}^{-1/2} - \Sigma_{jS}^{-1/2})\zeta_{jS}|$$
$$= \max_{jS} |(\widehat{\Sigma}_{jS}^{-1/2}/\Sigma_{jS}^{-1/2}) - 1| \max_{jS} |\Sigma_{jS}^{-1/2}\zeta_{jS}|$$

Applying the conditions above, the moderate deviations results for self-normalized sums are used to show (see [5] for detailed description of how to apply the results in [25] with the conditions above to get this result)

$$\max_{jS} |\Sigma_{jS}^{-1/2}\zeta_{jS}| = O_{P_n}\sqrt{N_n \log p}$$

Next we show that,

$$\max_{jS} |(\widehat{\Sigma}_{jS}^{-1/2}/\Sigma_{jS}^{-1/2}) - 1| = O_P\sqrt{N_n \log p/n}$$

This gives that the righthand term $\max_{jS} |(\widehat{\Sigma}_{jS}^{-1/2}/\Sigma_{jS}^{-1/2}) - 1| \max_{jS} |\Sigma_{jS}^{-1/2}\zeta_{jS}|$ vanishes asymptotically in probability after noting that $\frac{N_n^2 \log^2 p}{n} \to 0$ by the rate assumption.

Consider

$$\widehat{\Sigma}_{jS} - \Sigma_{jS} = \sum_{i=1}^{n} \breve{\psi}_{ijS}^2(\widehat{\epsilon}_{ijS}^2 - \epsilon_{ijS}{}^2)$$

$$\leqslant \sum_{i=1}^{n} \breve{\psi}_i^2[\psi_i([\theta_{jS}^*]_j - [\widehat{\theta})_{jS}]_j]^2 + 2\sum_{i=1}^{n} \breve{\psi}_{ijS}^2\epsilon_i\psi_{ijS}([\theta_{jS}^*]_j - [\widehat{\theta}_{jS}]_j)$$

Letting $d_{jS} = [\theta_{jS}]_j - [\widehat{\theta}_{jS}]_j$ then the above is bounded according to:

$$\leqslant \|d_{jS}\|_1^2 \sum_{i=1}^{n} \breve{\psi}_i^2\|\psi_i\|_\infty^2 + \|d_{jS}\|_1 \left\| \sum_{i=1}^{n} \breve{\psi}_{ijS}^2\epsilon_i\psi_{ijS} \right\|_\infty$$
$$\leqslant \|d_{jS}\|_1^2 O(n) + \|d_{jS}\|_1 O_P(\sqrt{N_n \log p})$$

We bound the quantity d_{jS} by

$$\max_{jS} \|d_{jS}\|_2 = \max_{jS} \|(\psi_{jS}'\psi_{jS})^{-1}\psi'\epsilon\|_2 \leqslant \sqrt{N_n} c_{\mathrm{irr}} \max_j |\frac{1}{n}\psi_j'\epsilon|$$

so that, $\max_{jS} \|[\theta_{jS}]_j - [\widehat{\theta}_{jS}]_j\|_1 \leqslant O(N_n) \max_j |\frac{1}{n}\psi_j'\epsilon|$.

Note that $\max_j |\frac{1}{n}\psi_j'\epsilon_i| \leqslant |\max_j \Sigma_{j\varnothing}^{-1/2}\frac{1}{n}\psi_j'\epsilon_i| \max_j \Sigma_{j\varnothing}^{1/2}$

Using Condition 3 and applying the theory for moderate deviation bounds for self-normalized sums enabled by Condition 4, this gives: $|\max_j \sqrt{n}\Sigma_{j\varnothing}^{-1/2}\frac{1}{n}\psi_j'\epsilon_i| \max_j \Sigma_{j\varnothing}^{1/2}/\sqrt{n} = O_P(\sqrt{\log p})O_P(1)$. This implies the desired bounds.

The final task is bounding the quantiles of $\Sigma_{jS}^{-1/2}\zeta_{jS}$. This is a self-normlized sum. The denominator has the form

$$= \sqrt{\sum_{i=1}^{n} \breve{\psi}_{ijS}^2({\epsilon_i}^2 + 2\epsilon_i \xi_{ijS} + {\xi_{ijS}}^2)}$$

which due to the large deviation assumption stated in Condition 4, is with high probability smaller than

$$\sqrt{\sum_{i=1}^{n} \breve{\psi}_{ijS}^2 \epsilon_i^2}$$

In the numerator of the self-normlized sum $\Sigma_{jS}^{-1/2}\zeta_{jS}$, we have

$$\breve{\psi}_{jS}'(\epsilon + \xi_{jS}) = \breve{\psi}_{jS}'\epsilon$$

from the fact that ξ_{jS} and $\breve{\psi}_{jS}$ are exactly orthogonal (using that fact that the covariates are fixed). Note that with random covariates, we would need to bound terms of the form $\frac{\sum_{i=1}^n \xi_i^* \psi_k(x_i)}{\sqrt{\sum_{i=1}^n \xi_i^* \psi_k(x_i)^2}}$ ranging over j, S.

Consider the event $\mathcal{R}_t$ where for each $k \leqslant p$

$$\frac{\sum_{i=1}^n \epsilon_i \psi_k(x_i)}{\sqrt{\sum_{i=1}^n \epsilon_i \psi_k(x_i)^2}} \leqslant t$$

Next note that on $\mathcal{R}$, the following sequence of inequalities holds

$$\left(\sum_{i=1}^{n} \sum_{k \in jS} \eta_k \psi_k(x_i)\epsilon_i\right)^2 \leqslant \left(t \sum_{k \in jS} \eta_k \sqrt{\sum_{i=1}^{n} \psi_k(x_i)^2 \epsilon_i^2}\right)^2$$

Next, define the matrix Ψ_{jS}^{ϵ} such that $[\Psi_{jS}^{\epsilon}]_{k,l} = \sum_{i=1}^n \epsilon_i^2 \psi_k(x_i)\psi_l(x_i)$ for $k, l \in jS$. Similarly, define Ψ_{jS} such that $[\Psi_{jS}]_{k,l} = \sum_{i=1}^n \psi_k(x_i)\psi_l(x_i)$. With this definition we have

$$\left(\sum_{i=1}^{n} \sum_{k \in jS} \eta_k \psi_k(x_i)\epsilon_i\right)^2 \leqslant {\tau_{jS}}^2 t^2 \eta' \Psi^{\epsilon} \eta = {\tau_{jS}}^2 t^2 \sum_{i=1}^{n} \left(\sum_{k \in jS} \eta_k \psi_k(x_i)\right)^2 \epsilon_i^2$$

So that

$$|\Sigma_{jS}^{-1/2}\zeta_{jS}| \leqslant \tau_{jS} t \ \text{ on } \mathcal{R}_t$$

Unfortunately, the quantity τ is infeasible since it involves ϵ_i terms. Note that in constructing testing thresholds, we had proposed replacing

Ψ^{ϵ} with the analogously defined estimate $\Psi^{\widehat{\epsilon}}$ (defined so that $[\Psi^{\epsilon}_{jS}]_{k,l} = \sum_{i=1}^n \epsilon_i^2 \psi_k(x_i)\psi_l(x_i)$ for $k,l \in jS$.) Under calculations like before we have

$$\max_{j,|S|\leqslant N} \|\Psi^{\epsilon}_{jS} - \Psi^{\widehat{\epsilon}}_{jS}\|_{2\to 2} \to_P 0$$

which implies that uniformly over $j, |S| \leqslant N_n$, $\widehat{\tau}_{jS} - \tau_{jS} \to_P 0$.

Let $t_\alpha := \Phi^{-1}(1-\alpha/p)$. Then by construction,

$$T_{jS\alpha} = 1 \iff |\widehat{V}_{jS}^{-1/2}\widehat{\theta}| \geqslant c_\tau \widehat{\tau}_{jS} t_\alpha$$

Note that, as argued above using moderate deviation bounds applied by Condition Ex1.4, we have $P(\mathcal{R}_{t_\alpha}) = \alpha + o(1)$. By the above, with probability $1-\alpha+o(1)$,

$$|\widehat{V}_{jS}^{-1/2}([\widehat{\theta}_{jS}]_j - [\theta^*_{jS}]_j)| \leqslant \tau_{jS} t_\alpha + o(1)$$

The above two inequalities imply that whenever $T_{jS\alpha} = 1$,

$$|\widehat{V}_{jS}^{-1/2}[\theta^*_{jS}]_j| \geqslant (c_\tau \widehat{\tau}_{jS} - \tau_{jS}) t_\alpha - o(1)$$

Also, with probability $1-o(1)$, for n sufficiently large,

$$\widehat{V}^{1/2}(c_\tau \widehat{\tau}_{jS} - \tau_{jS}) t_\alpha \geqslant V_{jS}^{1/2} \frac{c_\tau + 1}{2} \tau_{jS} t_\alpha.$$

Summarizing gives that with probability $1-\alpha-o(1)$:

$$\left\{ T_{jS\alpha} = 1 \implies |[\theta^*_{jS}]_j| \geqslant V_{jS}^{1/2} \frac{c_\tau + 1}{2} \tau_{jS} t_\alpha \right\}.$$

which is equivalent to

$$\left\{ |[\theta^*_{jS}]_j| \leqslant \frac{c_\tau + 1}{2} V_{jS}^{1/2} \tau t_\alpha \implies T_{jS\alpha} = 0 \right\}.$$

By similar logic, we have with probability $1-o(1)-\alpha$ the event:

$$\left\{ |[\theta^*_{jS}]_j| \geqslant (c_\tau + 1) V_{jS}^{1/2} \tau_{jS} t_\alpha \implies T_{jS\alpha} = 1 \right\}.$$

At this point, we point out that by assumption, $V_{jS}^{1/2} \times \sqrt{n}$ is with high probability bounded away from zero and above, for all j, S, by constants which are independent of n. The same is true for τ. Finally, $t_\alpha/\sqrt{\log(\alpha/p)} \to 1$. Qualitatively, these calculations can be used to show that $\{|[\theta^*_{jS}]_j| \leqslant O(\sqrt{\frac{\log p/\alpha}{n}}) \implies T_{jS\alpha} = 0\}$ and $\{|[\theta^*_{jS}]_j| \geqslant$

$O(\sqrt{\frac{\log p/\alpha}{n}}) \implies T_{jS\alpha} = 1\}$ with high probability. These facts are used in verifying Condition 2(I) and Condtion 2(II) when applying Theorem 1 to this problem.

Now suppose $T_{jS\alpha} = T_{kS\alpha} = 1$ and that $W_{jS} \geqslant W_{kS}$. We derive some facts which are useful for verifying Condition 2(III) for applying Theorem 1 to this problem. We have,

$$|\widehat{V}_{jS}^{-1/2}([\widehat{\theta}_{jS}]_j - [\theta^*_{jS}]_j) + \widehat{V}_{jS}^{-1/2}[\theta^*_{jS}]_j| \geqslant |\widehat{V}_{kS}^{-1/2}([\widehat{\theta}_{kS}]_k - [\theta^*_{kS}]_k) + \widehat{V}_{kS}^{-1/2}[\theta^*_{kS}]_k|$$

We lower bound the right hand side and upper bound the left hand side of the above inequality. We start with the right hand side. As above, $|V_{kS}^{-1/2}[\theta^*_{kS}]_k| \geqslant \frac{c_\tau+1}{2}\tau_{jS}t_\alpha$ and $|\widehat{V}_{kS}^{-1/2}([\widehat{\theta}_{kS}]_k - [\theta^*_{kS}]_k)| \leqslant \tau_{jS}t_\alpha$ imply that

$$W_{kS} \geqslant \frac{c_\tau - 1}{2}|\widehat{V}_{kS}^{-1/2}[\theta^*_{kS}]_k|$$

A similar argument shows that

$$\frac{c_\tau + 1}{2}|\widehat{V}_{jS}^{-1/2}[\theta^*_{jS}]_j| \geqslant W_{jS}$$

letting $\widehat{F}_{jkS} = \frac{A_{jS}\widehat{V}_{jS}^{-1/2}}{A_{kS}\widehat{V}_{kS}^{-1/2}}$, we have from our formula for $\Delta_j \mathcal{E}(S)$ above that

$$-\Delta_j \mathcal{E}(S) \geqslant \widehat{F}_{jkS}\frac{c_\tau - 1}{c_\tau + 1}(-\Delta_k \mathcal{E}(S))$$

Finally, $\widehat{F}_{jkS} \geqslant c$ with probability $1 - o(1)$.

As above, $\widehat{V}_{jS}^{1/2} = A_{jS}\widehat{\Sigma}_{jS}^{1/2}$ and $\widehat{\Sigma}_{jS}^{1/2} = \Sigma_{jS}^{1/2}(1 + o_P(1))$. Since Σ_{jS} is bounded in probability away from zero and above uniformly in $j, |S| \leqslant N_n$ and A_{jS} is similarly bounded away from zero and above uniformly. Therefore, there is a constant, suggestively c_2'' which is independent of n such that with probability $1 - o(1) - \alpha$:

$$-\Delta_j \mathcal{E}(S) \geqslant c_2'' \times (-\Delta_k \mathcal{E}(S)) \ \ \forall j, k, |S| \leqslant N_n : T_{jS\alpha} = T_{kS\alpha} = 1, W_{jS} \geqslant W_{kS}.$$

All these calculations verify key properties of the testing procedure. With these at hand, we apply Theorem 1. The bounds $s \leqslant O(1)\widehat{s}$ follows from $N_n/s \to \infty$ and $c_3(N_n) = O(1)$, $C_1 = O(1)$ and $C_2 = O(1)$, which follow from the above derivations. The bound on the prediction errors follows similarly. This completes the proof of Theorem 2.