

Kozbur, Damian

Working Paper

Testing-based forward model selection

Working Paper, No. 283

Provided in Cooperation with:

Department of Economics, University of Zurich

Suggested Citation: Kozbur, Damian (2018) : Testing-based forward model selection, Working Paper, No. 283, University of Zurich, Department of Economics, Zurich, <https://doi.org/10.5167/uzh-151160>

This Version is available at:

<https://hdl.handle.net/10419/192892>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



**University of
Zurich** ^{UZH}

University of Zurich
Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 283

Testing-Based Forward Model Selection

Damian Kozbur

First version: January 2015
This version: April 2018

Testing-Based Forward Model Selection

Damian Kozbur

University of Zürich

Department of Economics

Schönberggasse 1, 8001 Zürich

e-mail: damian.kozbur@econ.uzh.ch.

Abstract: This paper introduces and analyzes a procedure called *Testing-Based Forward Model Selection* (TBFMS) in linear regression problems. This procedure inductively selects covariates that add predictive power into a working statistical model before estimating a final regression. The criterion for deciding which covariate to include next and when to stop including covariates is derived from a profile of traditional statistical hypothesis tests. This paper proves probabilistic bounds for prediction error and the number of selected covariates, which depend on the quality of the tests. The bounds are then specialized to a case with heteroskedastic data with tests derived from Huber-Eicker-White standard errors. TBFMS performance is compared to Lasso and Post-Lasso in simulation studies. TBFMS is then analyzed as a component into larger post-model selection estimation problems for structural economic parameters. Finally, TBFMS is used to illustrate an empirical application to estimating determinants of economic growth.

MSC 2010 subject classifications: 62J05, 62J07, 62L12.

JEL subject classifications: C55.

Keywords and phrases: model selection, forward regression, sparsity, hypothesis testing.

*The content of this paper draws from two separate papers posted on ArXiv (<https://arxiv.org/abs/1702.01000> ([42]) and <https://arxiv.org/abs/1512.02666> (the current paper)), which were originally created as two different projects. The two projects have now been merged in preparation for the publication process. No original material is submitted simultaneously to multiple peer-reviewed journals. A mechanical statement of a special case of TBFMS appears in [43], which is published in the Papers and Proceedings issue of the American Economic Review, which is not peer reviewed. That version neither claims nor derives any theoretical results.

* First version: Jan 2015. This version April 17, 2018. I gratefully acknowledge helpful discussion with Christian Hansen, Tim Conley, Dan Zou, 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 from UZH and from the ETH Fellowship program.

1. Introduction

This paper considers a procedure called *Testing-Based Forward Model Selection* (TBFMS) for high-dimensional econometric problems, which are estimation problems characterized by settings in which the number of observed characteristics per observation in the data is large.¹ High-dimensional econometrics is a leading area of current research because of recent rapid growth in data availability and computing capacity coupled with the important need to extract as much useful information from data in a way that allows precise and rigorous testing of scientific hypotheses. Working within the flexibility of a high-dimensional framework allows researchers to fully exploit richer data sets both in prediction problems and in structural inference problems.

The primary settings of this paper are high-dimensional sparse linear regression models, in which the number of covariates is allowed to exceed the sample size. A key challenge with a high-dimensional data set is that estimation requires dimension reduction or regularization to avoid statistical overfitting. One regularizing structure used often in the recent econometrics and statistics literature is *sparsity*. A sparsity assumption imposes that the regression function relating the outcome and the covariates can be approximated by a regression of the outcome on a small, ex ante unknown subset of covariates. Under sparsity, there are several consistent estimation procedures (further reviewed below) that work by enforcing that the estimated regression function be sparse or small under an appropriate norm.

An appealing class of techniques for high-dimensional regression problems are *Greedy* algorithms. These are procedures that inductively select individual covariates into a working model (i.e., a collection of covariates) until a stopping criterion is met. A linear regression restricted to the final selected model is then estimated. A leading example is *Simple Forward Selection*,² which chooses the covariate that gives the highest increase of in-sample R-squared above the previous working model. This class of techniques is widely used because they are intuitive and simple to implement.

In practice, deciding which covariate gives the best additional predictive power is com-

¹High-dimensional data may arise in several ways. The data may be intrinsically high-dimensional with many characteristics per observation. Alternatively, even with a relatively small number of covariates, researchers may still be obtain a large final set of potential covariates formed by interactions and transformations of underlying covariates.

²Simple Forward Selection is not standard nomenclature, but is used here in order to have a parallel language with Testing-Based Forward Model Selection. The literature is varied and uses several names including Forward Regression and Forward Stepwise Regression.

plicated 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 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 the introduction of a testing-based strategy. Mechanically, TBFMS begins with an empty model. The procedure then tests whether any covariates provide additional predictive capability in the population. The selection stops when no tests return a significant covariate. Selection into the model is then based on the largest value of an associated test statistic.

Though the prior literature has not analyzed greedy algorithms that explicitly incorporate hypothesis testing, there are several earlier analyses of Simple Forward Selection.^{3,4} [68] gives bounds on the performance and number of selected covariates under a β -min condition which requires the minimum magnitude of non-zero coefficients to be suitably bounded away from zero. [72] and [64] prove performance bounds for greedy algorithms under a strong irrepresentability condition, which restricts the empirical covariance matrix of the predictors. [27] 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 addition to Simple Forward Selection, there are several related procedures in which estimation is done in stages. These include a method that is not strictly greedy called *Forward-Backward Selection*, which proceeds similarly to Simple Forward Selection but allows previously selected covariates to be discarded from the working model at certain steps (see [72]). Another related class of methods are called boosting methods. Boosting methods inductively select covariates predictive of a

³A mechanical statement of TBFMS in a specific case with heteroskedastic data is given in [43], which is a Papers and Proceedings publication by the current author and is not peer-reviewed. No derivations or theoretical properties are stated or claimed there.

⁴TBFMS using different tests than proposed here is natively programmed in some statistical software, including SPSS, but is not previously formally justified.

linear combination of estimated residuals and the outcome at each step (among many other references, see [17], see also [48] for additional results and applications in econometrics.)

As a preliminary, (before proceeding to the analysis of TBFMS), this paper proves new bounds on the predictive performance and number of selected covariates for Simple Forward Selection. The conditions required here are weaker than those used in [72] and [64] and impose no β -min restrictions or irreducibility. The convergence rates here are most similar to the analysis of a Forward-Backward Selection in [73], but require markedly different analysis since there is no chance to correct “over-selection mistakes.”

The analysis of Simple Forward Selection lays the groundwork for deriving statistical performance bounds for TBFMS. This paper derives performance bounds for TBFMS which depend directly on the quality of the profile of tests considered, as measured by 5 constants which characterize size and power. The abstract results for TBFMS are used to derive asymptotic bounds for various sequences of data-generating processes. As an example, concrete tests for heteroskedastic data constructed from Huber-Eicker-White standard errors are used to construct t -tests and explicit rates of convergence are calculated.

This paper complements an emerging branch of literature on sequential testing (see [31], [45], [63], [28]). This literature considers hypothesis testing in stages, in which tests in later stages can depend on testing outcomes in earlier stages. In various settings, properties like family-wise error rates of proposed testing procedures can be controlled sequences of hypothesis tests. In all cases, the authors note that the testing procedures are complementary to forward model selection problems as they guide which variables should be selected and offer principled stopping rules. The interest in the current paper lies primarily in the statistical properties and performance bounds of estimates and fits based on a selected model from a greedy algorithm that leverages testing. Future work may potentially combine the two types of problems.

There are many other sensible 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 [29] [62], [35] [21] [4], [5], [15], [19], [18] [20], [21], [36], [39], [40], [46], [47], [50], [56], [62], [65], [67], [70], [7], [16], [7], among many more. The asymptotic estimation rates calculated for TBFMS here

match those standard for Lasso and Post-Lasso.

After developing several theoretical bounds, a simulation study illustrates the relative performance of TBFMS to Lasso and Post-Lasso regression. The simulation study shows that there are data-generating processes under which forward selection outperforms Lasso regression in terms of prediction and estimation error.

In economic applications, models learned using formal model selection are often used in subsequent estimation steps with the final goal of learning a structural parameter of interest. One example is the selection of instrumental variables for later use in a first-stage regression (see [6], [34]). Another example is the selection of a conditioning set, to properly control for omitted variables bias when there are many control variables (see [14], [66], [11], [41]). 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. Such applications require a model selection procedure with a hybrid objective: (1) produce a good fit, and (2) return a sparse set of variables. This paper addresses both objectives, and therefore provides adequately tight bounds using strictly forward selection for application in causal post-estimation analysis.

Finally, TBFMS is illustrated in an economic application. The application revisits 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, the set of possible controls for geography is expanded to 16. Since 16 is not vanishingly small relative to the sample size of 64, this example lies within the space of high-dimensional estimation problems. The most relevant geographic controls are then chosen using TBFMS. To be robust to model selection mistakes and not suffer classical problems known to be associated with pretesting, three model selection steps are required (see [12], [14]), each separately applying TBFMS. These are: (1) select those geographic variables predictive of output; (2) select those geographic controls predictive of quality of institution; (3) select those geographic controls predictive of European settler mortality. Final estimates of the effects of institution quality on growth are generated with standard

IV estimation applied using the union of selected controls in the three selection steps. The 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. Precursor: Sharp Convergence Rates for Simple Forward Selection without β -min or Irrepresentability Conditions

This section proves a precursory result about Simple Forward Selection which is new in the high-dimensional econometrics and statistics literature despite the literature's maturity. The procedure is defined formally below and is similar to TBFMS, but uses a single threshold rather than a profile of hypothesis tests in determining the selection of covariates. The framework set out in this section is also helpful in terms of outlining minimal structure needed to facilitate the method of analysis in the formal arguments that follow.⁵

2.1. Framework

A realization of data of sample size n is given by $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ and is generated by a joint distribution P . The data consists of a set of covariates $x_i \in \mathbb{R}^p$, as well as outcome variables $y_i \in \mathbb{R}$ for each observation $i = 1, \dots, n$. The data satisfy

$$y_i = x_i' \theta_0 + \varepsilon_i$$

for some unknown parameter of interest $\theta_0 \in \mathbb{R}^p$ and unobserved disturbance terms $\varepsilon_i \in \mathbb{R}$. The parameter θ_0 is sparse in the sense that the set of non-zero components of θ_0 , denoted $S_0 = \text{supp}(\theta_0)$, has cardinality $s_0 < n$.

Define a loss function $\ell(\theta)$

$$\ell(\theta) = \mathbb{E}_n[(y_i - x_i' \theta)^2]$$

where $\mathbb{E}_n[\cdot] = \frac{1}{n} \sum_{i=1}^n (\cdot)$ denotes empirical expectation. Note that $\ell(\theta)$ depends on \mathcal{D}_n , but this dependence is suppressed from the notation. Define also

$$\ell(S) = \min_{\theta: \text{supp}(\theta) \subseteq S} \ell(\theta).$$

⁵This section draws material from the draft [42], also written by the current author, which was originally a separate project and posted on ArXiv, but is now merged into the current paper in preparation for the publication process.

The estimation strategy proceeds by first searching for a sparse subset $\widehat{S} \subseteq \{1, \dots, p\}$, with cardinality \widehat{s} , that assumes a small value of $\ell(S)$, followed by estimating θ_0 with least squares via

$$\widehat{\theta} \in \arg \min_{\theta: \text{supp}(\theta) \subseteq \widehat{S}} \ell(\theta).$$

This gives the construction of the estimates $x'_i \widehat{\theta}$ for $i = 1, \dots, n$.

The set \widehat{S} is selected as follows. For any S define the incremental loss from the j th covariate by

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

Consider the greedy algorithm, which inductively selects the j th covariate to enter a working model if $-\Delta_j \ell(S)$ exceeds a threshold t :

$$-\Delta_j \ell(S) > t$$

and $\Delta_j \ell(S) \geq \Delta_k \ell(S)$ for each $k \neq j$. The threshold t is chosen by the user; it is the only tuning parameter required. This procedure is summarized formally here:

Algorithm 1. *Simple Forward Regression*

Initialize. Set $\widehat{S} = \emptyset$

For $1 \leq k \leq p$

If $-\Delta_j \ell(S) > t$ for some $j \in \{1, \dots, p\} \setminus \widehat{S}$

Set $\widehat{j} \in \arg \max \{-\Delta_j \ell(S) : -\Delta_j \ell(S) > t\}$

Update $\widehat{S} = \widehat{S} \cup \{\widehat{j}\}$

Else

Break

Set $\widehat{\theta} \in \arg \min_{\theta: \text{supp}(\theta) \subseteq \widehat{S}} \ell(\theta)$.

2.2. Formal Analysis

In order to analyze Algorithm 1 and state the first theorem, a few more definitions are convenient. Define the empirical Gram matrix G by $G = \mathbb{E}_n[x_i x'_i]$. Let $\varphi_{\min}(s)(G)$ denote the minimal s -sparse eigenvalues given by

$$\varphi_{\min}(s)(G) = \min_{S \subseteq \{1, \dots, p\}: |S| \leq s} \lambda_{\min}(G_S)$$

where G_S is the principal submatrix of G corresponding to the component set S . The maximal sparse eigenvalues $\varphi_{\max}(s)(G)$ are defined analogously. Let

$$c_F(\widehat{s}) = (\widehat{s} + s_0)^{1/2} \varphi_{\min}(\widehat{s} + s_0)(G)^{-1} \left[2 \|\mathbb{E}_n[\varepsilon_i x'_i]\|_{\infty} + t^{1/2} \right].$$

Finally, for each positive integer m , let

$$c'_F(m) = 1 + 72 \times 1.783^2 \times \varphi_{\min}(m + s_0)(G)^{-5}.$$

Theorem 1. *Consider a data set \mathcal{D}_n of fixed sample size n with parameter θ_0 . Suppose the normalizations $\mathbb{E}_n[x_{ij}^2] = 1$ hold for each $j \leq p$. Then under Algorithm 1 with threshold t ,*

$$\mathbb{E}_n[(x'_i\theta_0 - x'_i\hat{\theta})^2]^{1/2} \leq c_F(\hat{s}).$$

In addition, for every integer $m \geq 0$ such that $t^{1/2} \geq 2\varphi_{\min}(m + s_0)(G)^{-1}\|\mathbb{E}_n[x_i\varepsilon_i]\|_\infty$ and $m \leq |\hat{S} \setminus S_0|$, it holds that

$$m \leq c'_F(m)s_0.$$

The above theorem calculates explicit finite sample constants bounding the prediction error norm. The second statement is a tool for bounding the number of selected covariates. In particular, setting $m^* = \min\{m : m > c'_F(m)s_0\}$ implies that

$$\hat{s} \leq m^* + s_0$$

provided $t^{1/2} \geq 2\varphi_{\min}(m^* + s_0)(G)^{-1}\|\mathbb{E}_n[x_i\varepsilon_i]\|_\infty$.

The theorem is proven in detail in the Appendix. The first statement of Theorem 1 is proven through a combination of two techniques. The first technique creates an analogue of the *Basic Inequality* from standard Lasso analysis⁶ that states that $\ell(\hat{S} \cup S_0) \leq \ell(\theta_0)$. The second technique uses a Lemma from [27] which bounds $\ell(\hat{S}) - \ell(\hat{S} \cup S_0)$ in terms of s_0, t , and the sparse eigenvalues of G . Together, these produce a bound for $\ell(\hat{S}) - \ell(\theta_0)$ that translates to the statement in Theorem 1. The proof of the second statement of the theorem involves bounding \hat{s} . In the analysis of Lasso-based estimation, bounds for \hat{s} are typically derived with the aid of KKT conditions for the solution to the Lasso problem. Because Simple Forward Regression is iterative, no such optimality conditions exist. Instead, the proof leverages a pigeonhole principle which is motivated by the following reasoning. \hat{s} being sufficiently large relative to s_0 forces the existence of some subset $S \subseteq \hat{S}$ such that $\lambda_{\min}(G_S)^{-1}$ is correspondingly large.⁷ This is possible because all covariates indexed in \hat{S} must exhibit some correlation to covariates indexed in S_0 . At the same time $\lambda_{\min}(G_S)^{-1}$ cannot exceed $\varphi_{\min}(\hat{s} + s_0)(G)^{-1}$. The desired bounds controlling \hat{s} are then deduced.

⁶For Lasso estimation with penalty level λ , the Basic Inequality asserts that $\ell(\hat{\theta}) + \lambda\|\hat{\theta}\|_1 \leq \ell(\theta_0) + \lambda\|\theta_0\|_1$.

⁷Formally, this pigeonhole principle is carried out by constructing an operator $L^\infty(\mathbb{R}^{s_0}) \rightarrow L^1(\mathbb{R}^{s_0})$ which depends on projections of covariates indexed in \hat{S} onto covariates indexed in S_0 , and then lower bounding its norm with the aid of Grothendieck's inequality ([32], see for a review, [53]; see also the exact form described in [33]; see the appendix for more details.)

The statement in Theorem 1 gives finite sample bounds which are completely deterministic in the sense that they hold for every possible realization of the data. Furthermore, the proof does not use the random nature of \mathcal{D}_n at any step. As a result, the bounds are very general, but it is helpful for interpretation to consider the convergence rates implied by Theorem 1 under asymptotic conditions on \mathcal{D}_n . Consider a sequence of random data sets $(\mathcal{D}_n)_{n \in \mathbb{N}}$ generated by joint distributions $\{P = P_n\}_{n \in \mathbb{N}}$. For each n , the data again satisfy $y_i = x_i' \theta_0 + \varepsilon_i$. In what follows, the parameters θ_0 , the thresholds t , distribution P , the dimension p of x_i , etc. can all change with n .

Condition 1 (*Asymptotic Regularity*). The sparsity satisfies $s_0 = o(n)$. There is a sequence K_n for which $s_0 = o(K_n)$ and there is a bound $\varphi_{\min}(K_n)(G)^{-1} = O(1)$ which holds with probability $1 - o(1)$. The normalizations $\mathbb{E}_n[x_{ij}^2] = 1$ hold a.s. for every $j \leq p$. The threshold satisfies a bound $t = O(\log p/n)$. In addition, $t^{1/2} \geq 2\varphi_{\min}(K_n)(G)^{-1} \|\mathbb{E}_n[x_i \varepsilon_i]\|_\infty$ with probability $1 - o(1)$.

The rates assumed in Condition 3 reflect typical rates achieved under various possible sets of low level conditions standard in the literature (ie. [6]). Condition 1 asserts three important statements. The first statement bounds the size of S_0 and requires that the sparsity level is small relative to the sample size. The second statement 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 G has full rank. In the high-dimensional setting, G will be singular if $p > n$ and may have an ill-behaved inverse even when $p \leq n$. However, good performance of many high-dimensional estimators only requires good behavior of certain moduli of continuity of G . There are multiple formalizations and moduli of continuity that can be considered here; see [15]. This analysis focuses on a simple eigenvalue condition which was used in [6]. Condition 2 could be shown to hold under more primitive conditions by adapting arguments found in [7] which build upon results in [70] and [58]; see also [57]. Condition 2 is notably weaker than previously used irrepresentability conditions. Irrepresentability conditions require that for certain sets S and $k \notin S$, letting x_{iS} be the subvector of x_i with components $j \in S$, that $\|\mathbb{E}_n[x_{iS} x'_{iS}]^{-1} \mathbb{E}_n[x_{iS} x'_{ik}]\|_1$ is strictly less than 1. The normalization $\mathbb{E}_n[x_{ij}^2] = 1$ is used to keep exposition concise and can be relaxed (and, e.g., is relaxed in Theorem 5).

The final statement in Condition 1 is a regularization condition similar to regularization conditions common in the analysis of Lasso. The condition, requires $t^{1/2}$ to dominate

a multiple of the $\|\mathbb{E}_n[x_i\varepsilon_i]\|_\infty$. This condition is stronger than that typically encountered with Lasso, because the multiple depends on the sparse eigenvalues of G . To illustrate why such a condition is useful, let \tilde{x}_{ij} denote x_{ij} residualized away from previously selected regressors and renormalized. Then even if $\mathbb{E}_n[x_{ij}\varepsilon_i] < t^{1/2}$, $\mathbb{E}_n[\tilde{x}_{ij}\varepsilon_i]$ can exceed $t^{1/2}$ resulting in more selections into the model. Nevertheless, using the multiple $2\varphi_{\min}(K_n)(G)^{-1}$ which stays bounded with n , is sufficient to ensure that \hat{s} does not grow faster than s_0 . Furthermore, this requirement does not implicitly impose a β -min condition and does not implicitly impose irrepresentability.

From a practical standpoint, Condition 1 does however implicitly require the user to know more about the design of the data in choosing an appropriate t . Choosing feasible thresholds which satisfy a similar condition to Condition 3 is considered in the next section where TBFMS is developed.

Theorem 2. *Consider a sequence of data sets \mathcal{D}_n indexed by n with parameters θ_0 and threshold t which satisfy Condition 1. Suppose $\hat{\theta}$ is obtained by Algorithm 1. Then there are bounds*

$$\mathbb{E}_n[(x'_i\theta_0 - x'_i\hat{\theta})^2]^{1/2} = O\left(\sqrt{\frac{s_0 \log p}{n}}\right),$$

$$\hat{s} \leq O(s_0).$$

which hold with probability $1 - o(1)$ as $n \rightarrow \infty$.

More explicitly, the implied O constants and o sequence in bounds for Theorem 2 are understood to depend only on the implied O constants and o sequences in Condition 1.

The theorem shows that Simple Forward Selection can obtain asymptotically the same convergence rates (specifically $\sqrt{s_0 \log p/n}$ for the quantities $\mathbb{E}_n[(x'_i\theta_0 - x'_i\hat{\theta})^2]^{1/2}$) as other high-dimensional estimators like Lasso, provided an appropriate threshold t is used. In addition, it selects a set with cardinality commensurate with s_0 .

Finally, two direct consequences of Theorem 2 are bounds on the deviations $\|\hat{\theta} - \theta_0\|_1$ and $\|\hat{\theta} - \theta_0\|_2$ of $\hat{\theta}$ from underlying unknown parameter θ_0 . Theorem 3 above shows that deviations of $\hat{\theta}$ from θ_0 also achieve rates typically encountered in high-dimensional estimators like Lasso.

Theorem 3. *Consider a sequence of data sets \mathcal{D}_n with parameters θ_0 and thresholds t which satisfy Condition 1. Suppose $\hat{\theta}$ is obtained by Algorithm 1. Then there are bounds*

$$\|\theta_0 - \hat{\theta}\|_2 = O\left(\sqrt{\frac{s_0 \log p}{n}}\right) \text{ and } \|\theta_0 - \hat{\theta}\|_1 = O\left(\sqrt{\frac{s_0^2 \log p}{n}}\right)$$

which hold with probability $1 - o(1)$ as $n \rightarrow \infty$.

3. Testing-Based Forward Model Selection

The previous section presented results on convergence rates Simple Forward Selection in a very simple context. The results of Theorems 1 are useful in terms of developing intuition and proof techniques for inductive variable selection algorithms. However, in terms of practical implementation, Section 2 leaves the question of how to choose a threshold unanswered. This section develops TBFMS in order to analyze feasible, data-driven ways to decide which covariates to select, and when to stop selecting.

3.1. Framework

The basic framework for this section is similar. Again, the observed data is given by $\mathcal{D}_n = (x_i, y_i)_{i=1}^n$ is generated by P and satisfies $y_i = x_i'\theta_0 + \varepsilon_i$ for a parameter θ_0 which is sparse with s_0 non-zero components supported on S_0 . Define $\ell(\theta)$ and $\ell(S)$ as before.

Define the expected loss function $\mathcal{E} : \mathbb{R}^p \rightarrow \mathbb{R}$ by

$$\mathcal{E}(\theta) = \mathbb{E} [\mathbb{E}_n(y_i - x_i'\theta)^2]$$

and note that $\mathcal{E}(\theta) = \mathbb{E}\ell(\theta)$. Extend the definition of \mathcal{E} to apply also as a map $\mathcal{E} : 2^{\{1, \dots, p\}} \rightarrow \mathbb{R}$ by $\mathcal{E}(S) = \min_{\theta: \text{supp}(\theta) \subseteq S} \mathcal{E}(\theta)$. Similarly to before, for any S define the incremental loss from the j th covariate by

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

Within the class of greedy algorithms, it would be preferable to consider a greedy algorithm which inductively selects the j th covariate to enter a working model if $\Delta_j \mathcal{E}(S)$ is large and $\Delta_j \mathcal{E}(S) \geq \Delta_k \mathcal{E}(S)$ for each $k \neq j$. However, because $\Delta_j \mathcal{E}(S)$ cannot generally be directly observed from the data, the idea that follows is to make use of statistical tests to gauge the magnitude of $\Delta_j \mathcal{E}(S)$. Consider a set of tests given by

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

Assume that the tests reject ($T_{jS\alpha} = 1$) for large values of a test statistic W_{jS} .

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, additional tests $T_{jkS\alpha}$ associated

to $H_0 : \Delta_j \mathcal{E}(S) \geq \Delta_k \mathcal{E}(S)$ could be devised to break ties. The test statistic approach is natural for breaking potential multi-way ties. To summarize, the algorithm for forward selection given the hypothesis tests $(T_{jS\alpha}, W_{jS})$ is now given formally.

Algorithm 2. *Testing-Based Forward Model Selection*

Initialize. Set $\widehat{S} = \emptyset$.

For $1 \leq k \leq p$:

If: $T_{j\widehat{S}\alpha} = 1$ for some $j \notin \widehat{S}$,

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

Update: $\widehat{S} = \widehat{S} \cup \{\widehat{j}\}$.

Else: *Break.*

Set: $\widehat{\theta} \in \arg \min_{\theta: \text{supp}(\theta) \subset \widehat{S}} \mathbb{E}_n(y_i - x_i' \theta)^2$.

3.2. Formal Analysis

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.

Condition 2 (*Hypothesis Tests*). There is an integer $K_{\text{test}} > s_0$ and constants $\alpha, \delta_{\text{test}}, c_{\text{test}}, c'_{\text{test}}, c''_{\text{test}} > 0$ such that each of the following conditions hold.

1. The tests have power in the sense that

$$\mathbb{P}(\{T_{jS\alpha} = 1 \text{ for every } j, |S| \leq K_{\text{test}} \text{ such that } -\Delta_j \mathcal{E}(S) \geq c_{\text{test}}\}) \geq 1 - \frac{1}{3} \delta_{\text{test}}.$$

2. The tests control size in the sense that

$$\mathbb{P}(\{T_{jS\alpha} = 1 \text{ for some } j, |S| \leq K_{\text{test}} \text{ such that } -\Delta_j \mathcal{E}(S) \leq c'_{\text{test}}\}) \leq \alpha + \frac{1}{3} \delta_{\text{test}}.$$

3. The tests are continuous in the sense that

$$\mathbb{P}(\{W_{jS} \geq W_{kS} \text{ for each } j, k, |S| \leq K_{\text{test}} \text{ such that } T_{jS\alpha} = T_{kS\alpha} = 1 \text{ and } -\Delta_j \mathcal{E}(S) \geq -c''_{\text{test}} \Delta_k \mathcal{E}(S)\}) \geq 1 - \frac{1}{3} \delta_{\text{test}}.$$

The constants c_{test} and c'_{test} measure quantities related to the size and power of the tests and provide a convenient language for subsequent discussion. The constant c''_{test} measures

the extent to which the test statistics W_{jS} reflect the actual magnitude of $\Delta_j \mathcal{E}(S)$. Note again that the hypothesis tests considered should not necessarily be thought of as providing a measure of statistical significance, but more precisely, they are simply a tool for model selection which coincidentally have many properties in common with traditional hypothesis tests.

Condition 3 (*Regularity*). Normalizations $\mathbb{E}[\mathbb{E}_n[x_{ij}^2]] = 1$ holds for all j . The residuals decompose into $\varepsilon_i = \varepsilon_i^o + \varepsilon_i^a$ where $\mathbb{E}[\mathbb{E}_n[\varepsilon_i^{o2}]] < \infty$, $\mathbb{E}[\mathbb{E}_n[\varepsilon_i^o x_{ij}]] = 0$ for all j , and $\mathbb{E}[\mathbb{E}_n[\varepsilon_i^{a2}]] \leq \frac{1}{2} \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1} c'_{\text{test}}$. Finally, $(2 + 1.783 \times 72 \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-5} c''_{\text{test}}{}^{-4}) s_0 < K_{\text{test}}$.

Condition 3 imposes regularity conditions for the class of models considered in the following theorem. First, ε_i is decomposed into an orthogonal component ε_i^o and an approximation component ε_i^a , each of which exhibits a different kind of regularity. The orthogonal component is orthogonal to the covariates in the population. The approximation component need not be orthogonal to the covariates, but its magnitude must be suitably controlled by the sparse eigenvalues of $\mathbb{E}[G]$ and by the parameter c'_{test} , which is a detection threshold for the profile of hypothesis tests $T_{jS\alpha}$. This decomposition allows for approximately sparse models similar to the framework of [6]. The fact that ε_i^a need not be orthogonal to the covariates also allows this framework to overlay onto many problems in traditional nonparametric econometrics.

Condition 3 also imposes conditions relating the sparse eigenvalues of $\mathbb{E}[G]$ with c''_{test} , s_0 , and K_{test} . Note that K_{test} measures the size of the set $S \subset \{1, \dots, p\}$ over which the hypothesis test perform well, as defined by Condition 2. Consequently, this condition requires that the hypothesis tests $T_{jS\alpha}$ perform sufficiently well over sets S , which must be larger when $\mathbb{E}[G]$ has small eigenvalues or when s_0 is large.

There are a few cases where Condition 3 can be simplified. Note that if $p > n$, even though the empirical Gram matrix is necessarily rank deficient, the population Gram matrix may be full rank. When $\mathbb{E}[G]$ is full rank, then the last statement of Condition 1 can be simplified to $(2 + 1.783 \times 72 \lambda_{\min}(\mathbb{E}[G])^{-5} c''_{\text{test}}{}^{-4}) s_0 < K_{\text{test}}$. In addition, the condition on ε_i^a implicitly imposes constraints on c'_{test} and $\varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1}$. When there is no approximation error, this requirement is no longer needed.

Let

$$\begin{aligned}
c_{\text{T}} &= s_0 \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1} c_{\text{test}} \\
c'_{\text{T}} &= 2 + 1.783 \times 72 \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-5} c''_{\text{test}}{}^{-4} \\
c''_{\text{T}}(\hat{s}) &= \varphi_{\max}(s_0 + \hat{s})(G)^{1/2} \varphi_{\min}(s_0 + \hat{s})(G)^{-1/2} \hat{s}^{1/2} \|\mathbb{E}_n[x_i \varepsilon_i]\|_{\infty} \\
&\quad + 3 \varphi_{\max}(s_0 + \hat{s})(G) \varphi_{\min}(\hat{s} + s_0)(G)^{-1/2} s_0^{1/2} c_{\text{test}}{}^{1/2} \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1}.
\end{aligned}$$

Theorem 4. Consider $\mathcal{D}_n \sim \mathbb{P}$ for some fixed n and $\{T_{jS\alpha}, W_{jS}\}$ such that Conditions 2 and 3 hold. Suppose $\hat{\theta}$ is obtained by Algorithm 2. Then the bounds

$$\begin{aligned}
\mathcal{E}(\hat{S}) - \mathcal{E}(S_0) &\leq c_{\text{T}} \\
\hat{s} &\leq c'_{\text{T}} s_0 \\
\mathbb{E}_n[(x'_i \theta_0 - x'_i \hat{\theta})^2] &\leq c''_{\text{T}}(\hat{s})
\end{aligned}$$

hold with probability at least $1 - \alpha - \delta_{\text{test}}$.

Theorem 4 is proven in the Appendix. It provides finite sample bounds on the performance of TBFMS. The outline for proving Theorem 4 is similar to that for proving Theorem 1. Theorem 4 works with the fact that $-\Delta_j \mathcal{E}(S) > c_{\text{test}}$ upon selection on the event implied by Condition 2 instead of the simpler $-\Delta_j \ell(S) > t$. In contrast to the proof of Theorem 1, the proof of Theorem 4 also addresses the possibility that if covariate j is selected ahead of covariate k , it is not necessarily the case that $-\Delta_j \mathcal{E}(S) > -\Delta_k \mathcal{E}(S)$. This is done by making use of the continuity constant c''_{test} in Condition 2.

Theorem 4 can be used to derive asymptotic estimation rates by allowing the constants to change with n . The next subsection provides an example to a linear model with heteroskedastic disturbances, where, under the stated regularity conditions, the prediction error and estimation error attain the rate $O_{\mathbb{P}}(s_0 \log p/n)$. These convergence rates therefore match typical Lasso and Post-Lasso rates.

Note that the results aim to control the hybrid objective, described in the introduction, of producing a good fit and returning a sparse set of variables. One useful implication of bounds controlling both \hat{s} and $\mathbb{E}_n[(x'_i \theta_0 - x'_i \hat{\theta})^2]$ is that the results can be applied to constructing uniformly valid post-model selection inference procedures (see [14]), in which for some applications, the prediction error bound alone is insufficient. Applications to inference are discussed in Section 4.

3.3. Example: Heteroskedastic Disturbances

This section gives an example of the use of Theorem 4 by illustrating an application of model selection in the presence of heteroskedasticity in the disturbance terms ε_i . The conditions required for the application of Theorem 4 are verified for a set of tests that are constructed based on the Heteroskedasticity-Consistent standard described in [69].

For shorthand, write x_{ijS} to be the vector with components x_{ik} with $k = j$ or $k \in S$. To construct the tests, begin with the least squares estimate of the regression y_i on x_{ijS} .

$$\hat{\theta}_{jS} = \mathbb{E}_n[x_{ijS}x'_{ijS}]^{-1}\mathbb{E}_n[x'_{ijS}y_i]$$

Define $\hat{\varepsilon}_{ijS} = y_i - x'_{ijS}\hat{\theta}_{jS}$. One heteroskedasticity robust estimate of the sampling variance of $\hat{\theta}_{jS}$, proposed in [69], is given by the expression

$$\hat{V}_{jS} = \frac{1}{n}\mathbb{E}_n[x_{ijS}x'_{ijS}]^{-1}\Psi_{jS}^{\hat{\varepsilon}}\mathbb{E}_n[x_{ijS}x'_{ijS}]^{-1}$$

where

$$\Psi_{jS}^{\hat{\varepsilon}} = \mathbb{E}_n[\hat{\varepsilon}_{ijS}^2 x_{ijS}x'_{ijS}].$$

Define the test statistics

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

Reject H_0 for large values of W_{jS} defined relative to an appropriately chosen threshold. To define the threshold, first let $\eta_{jS} := (1, -\beta'_{jS})'$ where β_{jS} is the coefficient vector from the least squares regression of $\{x_{ij}\}_{i=1}^n$ on $\{x_{ik}\}_{i=1, k \in S}^n$. Then define

$$\hat{\tau}_{jS} = \frac{\|\eta'_{jS}\text{Diag}(\Psi_{jS}^{\hat{\varepsilon}})^{1/2}\|_1}{\sqrt{\eta'_{jS}\Psi_{jS}^{\hat{\varepsilon}}\eta_{jS}}}.$$

The term $\hat{\tau}_{jS}$ will be helpful in addressing the fact that many different model selection paths are possible under different realizations of the data under P.⁸ Not taking this fact

⁸There is an unfortunate misprint in a Papers and Proceedings version of this paper, [43], in which the exponent 1/2 is missing from the term $\text{Diag}(\Psi_{jS}^{\hat{\varepsilon}})$. Note that [43] do not derive nor claim any theoretical properties, and instead just state the term $\hat{\tau}_{jS}$.

into account can potentially lead to false discoveries. The next condition states precisely the hypothesis tests $T_{jS\alpha}$.

Definition 1 (*Hypothesis Tests for Heteroskedastic Disturbances*). Let $c_\tau > 1$ and $\alpha > 0$ be parameters. Assign

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

The term $\Phi^{-1}(1 - \alpha/p)$ can be informally thought of as a Bonferroni correction term that takes into account the fact that there are p potential covariates. The term $c_\tau \hat{\tau}_{jS}$ can be informally thought of as a correction term that can account for the fact that the set S is random and can have many potential realizations. The simulation study uses the settings $c_\tau = 1.01$ and $\alpha = .05$.

Condition 5 (*Regularity for Data with Heteroskedasticity*). Consider a sequence of data sets $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n \sim \mathbb{P} = \mathbb{P}_n$. The observations (x_i, y_i) are i.n.i.d. across i and $y_i = x_i' \theta_0 + \varepsilon_i$ for some θ_0 with $s_0 = o(n)$. The residuals decompose into $\varepsilon_i = \varepsilon_i^o + \varepsilon_i^a$ such that a.s., $\mathbb{E}[\varepsilon_i^o | x_i] = 0$ and $\max_i |\varepsilon_i^a| = O(n^{-1/2})$. In addition, a.s., uniformly in i and n , $\mathbb{E}[\varepsilon_i^4 | x_i]$ are bounded above and $\mathbb{E}[\varepsilon_i^2 | x_i]$ is bounded away from zero. The covariates satisfy $\max_{j \leq p} \mathbb{E}_n[x_{ij}^2] = O(1)$ with probability $1 - o(1)$. There is a sequence K_n , where $s_0 = o(K_n)$, and bounds $\varphi_{\min}(K_n)(G)^{-1} = O(1)$, $\varphi_{\max}(K_n)(G)^{-1} = O(1)$, $\varphi_{\min}(K_n)(\mathbb{E}_n[(\varepsilon_i x_i)(\varepsilon_i x_i)'])^{-1} = O(1)$, and $\max_{|S| \leq K_n, j \notin S} \|\eta_{jS}\|_1 = O(1)$, which hold with probability $1 - o(1)$. The rate condition $K_n^4 \log^3 p/n = o(1)$ holds.

Condition 5, as before, gives conditions on the sparse eigenvalues, this time applying to both G and to $\mathbb{E}_n[(\varepsilon_i x_i)(\varepsilon_i x_i)']$. The requirement that $\varphi_{\min}(K_n)(G)^{-1}$ stay bounded can be relaxed, though the cost is that the proof of Theorem 5 could not directly call upon Theorem 4 (since c_{Γ}' involves a maximal sparse eigenvalue). In addition, Condition 5 assumes a bound on η_{jS} that may be strong in some cases. Previous results in [64], [72] assume the strict condition that $\max_{j \notin S_0} \|\eta_{jS_0}\|_1 < 1$, which is the genuine irrepresentability condition, in analysis of inductive variable selection algorithms. Here, the requirement < 1 is replaced by the weaker requirement $= O(1)$. Other authors, for instance [49], use conditions analogous to $\max_{|S| \leq K_n, j \notin S} \|\eta_{jS}\|_1 = O(1)$ in the context of learning high-dimensional

graphs, and note that the relaxed requirement is satisfied by a much broader class of data-generating processes. Analogous bounds on $\|\eta_{jS}\|_1$ were not required in Theorem 1, since the proof of Theorem 1 does not leverage bounds relating \widehat{W}_{jS} to the self-normalized sums $\mathbb{E}_n[x'_{ij}\varepsilon_i]/\sqrt{\mathbb{E}_n[x^2_{ij}\varepsilon_i^2]}$, $j \leq p$. Failure of the $O(1)$ bound would lead only to slightly slower convergence rates. Condition 5 also states regularity conditions on moments of ε_i and x_i , which are useful for proving laws of large numbers, central limit theorems, and moderate deviation bounds (see [38]). Finally, the rate condition assumes bounds on the relative sizes of s_0, p, n since $s_0 < K_n$.

Theorem 5. *Consider a sequence of data sets $\mathcal{D}_n \sim \mathbf{P} = \mathbf{P}_n$ which satisfies Condition 5. Suppose that $c_\tau > 1$ is fixed independent of n that $\alpha = o(1)$ with $n\alpha \rightarrow \infty$. Let $\widehat{\theta}$ be the estimate obtained from Algorithm 2 with tests defined by Definition 1. Then there are bounds*

$$\mathbb{E}_n[(x'_i\theta_0 - x'_i\widehat{\theta})^2]^{1/2} = O\left(\sqrt{\frac{s_0 \log(p/\alpha)}{n}}\right) \quad \text{and} \quad \widehat{s} = O(s_0)$$

which hold with probability at least $1 - \alpha - o(1)$ as $n \rightarrow \infty$.

The theorem is proven in the appendix by appealing to Theorem 4.

Analogous results potentially hold for dependent data using HAC-type standard errors (see [52], [3].) The required central limit results for such an application are beyond the scope of this work, though using the moderate deviation results for dependent data of [24] is a potential starting point. In addition, cluster-type standard errors for large- T -large- n and fixed- T -large- n panels can be used by adapting arguments from [11].

Allowing α to be fixed is possible under more restrictive conditions on the approximation error terms ε_i^a . Note that if $p > n$, then the rate $\log(p/\alpha)$ becomes equivalent to simply $\log(p)$.

As a final remark, note that there is a much simpler formulation for the hypothesis tests that ignores the $c_\tau \widehat{\tau}_{jS}$ terms. This results in the following definition.

Definition 2 (*Simplified Hypothesis Tests for Heteroskedastic Disturbances*). Let $c_\tau > 1$ and $\alpha > 0$ be parameters. Assign

$$T_{jS\alpha} = 1 \iff W_{js} \geq \Phi^{-1}(1 - \alpha/p).$$

These tests are based on a simple Bonferroni-type correction. Furthermore, though never previously formally justified, TBFMS using the simpler tests is natively programmed in some statistical software, including SPSS. It is unknown to the author at the time of this writing whether the same convergence rates can be attained using the simpler tests. This option is explored in some finite sample settings in the simulation study that follows. Evidence from the simulation study suggests that this option performs better than the more complex tests defined in Definition 1.

Another way to potentially lower the significance thresholds is to consider generalized error rates. The conditions set forth in Condition 2 require control of a notion resembling family-wise error rate uniformly over hypothesis tests $H_0 : \Delta_j \mathcal{E}(S)$ for $j \leq p$ and $|S| < K_{\text{test}}$ for some integer K_{test} . Other types of error rates like k -family wise error rate, false discovery rate, or false discovery proportion are potentially possible as well. In particular, the arguments in the proof of Theorem 4 would continue to be compatible with procedures that controlled an appropriate notion of false discovery proportion. In order to keep exposition concise, these extensions are not considered here.

3.4. TBFMS *Simulation*

The results in the previous sections suggest that estimation with TBFMS should produce good results in large sample sizes. This section simulates several different data-generating processes to evaluate the finite sample performance of TBFMS relative to select other procedures commonly used in high-dimensional regression problems.

This simulation study draws samples from the data-generating processes in Table 1.

Simulations are conducted with parameter settings $s_0 = 6$, $b_0 \in \{-0.5, 0.5\}$, $\rho_0 \in \{0, 0.5\}$, and $n \in \{100, 200, 300, 400, 500\}$ according to the data-generating process displayed in Table 1. The parameter s_0 dictates the sparsity. The parameter b_0 controls the nature of the coefficient vector. When $b_0 = -0.5$, the coefficients θ_{0j} alternate sign in j , and when $b_0 = 0.5$, the coefficients are all positive. The two different settings for b_0 create different interplay between the Toeplitz correlation structure in the covariates and their corresponding coefficients. The parameter ρ_0 controls the presence of heteroskedasticity in the disturbance terms ε_i . The terms ε_i are homoskedastic when $\rho_0 = 0$ and heteroskedastic otherwise. Finally, each setting of s_0 , b_0 , and ρ_0 is simulated in sample sizes ranging between 100 and 500. The dimensionality is always taken to be double the sample size so that $p = 2n$.

TABLE 1
TBFMS *Simulation Design*

Data:	$\mathcal{D}_n = (y_i, x_i, z_i)_{i=1}^n \text{ iid}$
DGP:	$y_i = x_i \theta_0 + \varepsilon_i$
	$p = \dim(x_i) = 2n, \quad \theta_{0j} = b_0^{j-1} \mathbf{1}_{j \leq s_0}$
	$x_{ij} \sim N(0, 1), \text{ with } \text{corr}(x_{ij}, x_{ik}) = 0.5^{ j-k }$
	$\varepsilon_i \sim \sigma_i N(0, 1), \quad \sigma_i = \exp(\rho_0 \sum_{j=1}^p 0.75^{(p-j)} x_{ij})$
Settings:	$s_0 = 6$
	$b_0 \in \{-0.5, 0.5\}$
	$\rho_0 \in \{0, 0.5\}$
	$n \in \{100, 200, 300, 400, 500\}$

This allows a visualization of the consistency properties of the displayed estimators in a high-dimensional asymptotic frame. Alternative parameterizations, for example pinning down the signal-to-noise ratio, are also interesting and possible, but are not displayed here in favor of brevity. Each simulation design is replicated 1000 times.

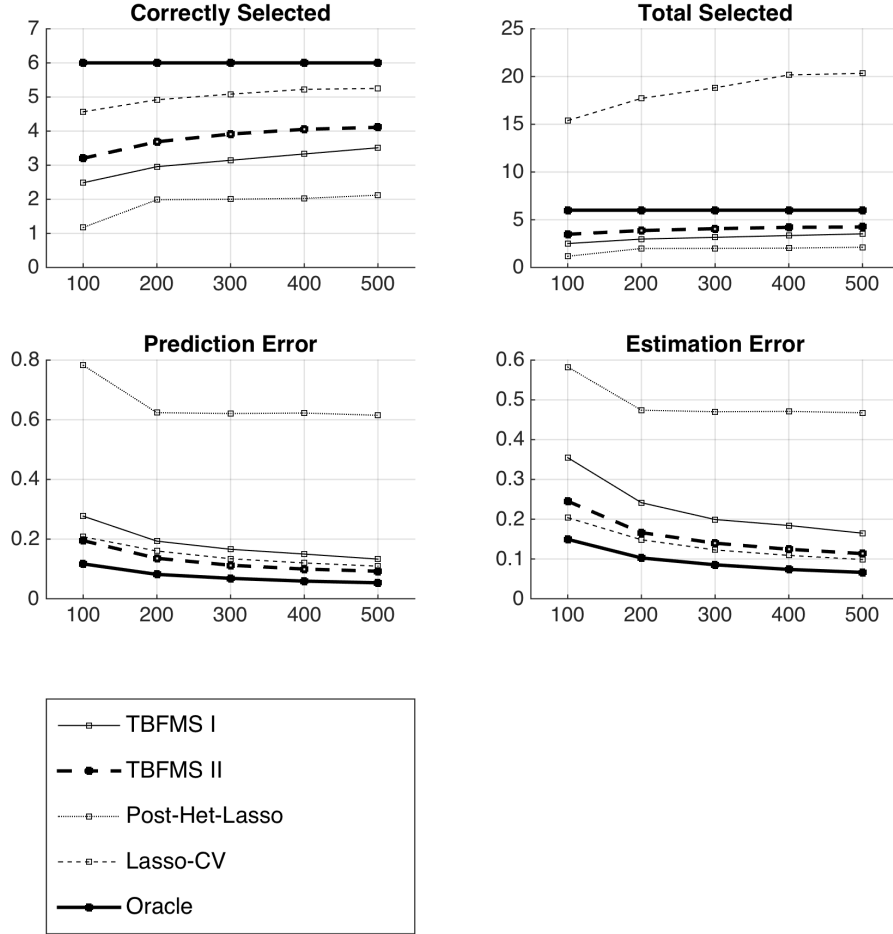
The simulation study compares five different estimators. In addition to two TBFMS estimators, two Lasso-based estimators and an infeasible estimator that knows the true support S_0 are presented.

1. **TBFMS I.** Implements Algorithm 2 with tests defined in Definition 1 with parameters $c_\tau = 1.01$, $\alpha = .05$.
2. **TBFMS II.** Implements Algorithm 2 with the simplified tests defined in Definition 2 with parameters $c_\tau = 1.01$, $\alpha = .05$.
3. **Lasso-CV.** Uses the original formulation of Lasso, implemented using glmnet, with penalty parameter chosen using 10-fold cross validation. $\hat{\theta}$ is the minimizer of the Lasso objective function and is not refit on the selected model.
4. **Post-Het-Lasso.** Uses the implementation found in [6], which is designed specifically to handle heteroskedastic disturbances. [6] requires two tuning parameters that are directly analogous to c_τ and α . These are set to $c_\tau = 1.01$ and $\alpha = .05$. $\hat{\theta}$ is refit on the selected model.
5. **Oracle** selects the model consisting of $\{j : [\theta_0]_j \neq 0\}$ and estimates a subsequent least squares regression.

The simulation results are presented in Figures 1–4. Each figure contains four plots that track various measures of estimation quality for the five estimators for fixed s_0, b_0 , and ρ_0 , and for n varying along the horizontal axis. The upper left plots display the number of correctly identified covariates from S_0 , averaged over simulation replications. The upper right plots display the total number of selected covariates, averaged over simulation replications. The bottom left plots display the prediction error defined as $\mathbb{E}_n[(x'_i\theta_0 - x'_i\hat{\theta})^2]^{1/2}$, averaged over simulation replications. The bottom right plots display estimation error defined as $\|\hat{\theta}_2 - \theta_0\|_2$, averaged over simulation replications.

In the simulations results, there is no single feasible estimator that dominates in every setting in terms of estimation error or prediction error. However, in all settings, Lasso-CV selects the most covariates (both in absolute terms and in terms of the number of correctly identified covariates), followed by TBFMS I, TBFMS II, and Post-Het-Lasso. There are instead important instances when each estimator performs better. In both settings with positive coefficients ($b_0 = 0.5$), Lasso-CV achieves the smallest estimation error with TBFMS I and TBFMS II having slightly higher estimation error. In these settings, however, the prediction error is smallest with TBFMS II. With alternative coefficients ($b_0 = -0.5$), however, TBFMS I and TBFMS II dominate Lasso-CV and Post-Het-Lasso on prediction error and estimation error. This suggests that the performance of these estimators depends on the configuration of the signal relative to the correlation structure of the covariates. Finally, the relative difference in performance across estimators is larger in the presence of heteroskedasticity. In the presence of heteroskedasticity, the Post-Het-Lasso exhibits faster improvement in estimation error and prediction error with increasing n , though it is still dominated by the other estimators. Note that each of the techniques, TBFMS I and Post-Het-Lasso, are theoretically valid for sequences of data-generating processes with heteroskedasticity. In addition, the properties of cross-validation with Lasso are only beginning to be understood (see [26] for analysis of lasso with cross-validation). But it is clear from this simulation study that Lasso-CV leads to selection of substantially more covariates to the extent that the effects of heteroskedasticity on the performance of the estimator are still not fully clear.

FIG 1. Simulation Results

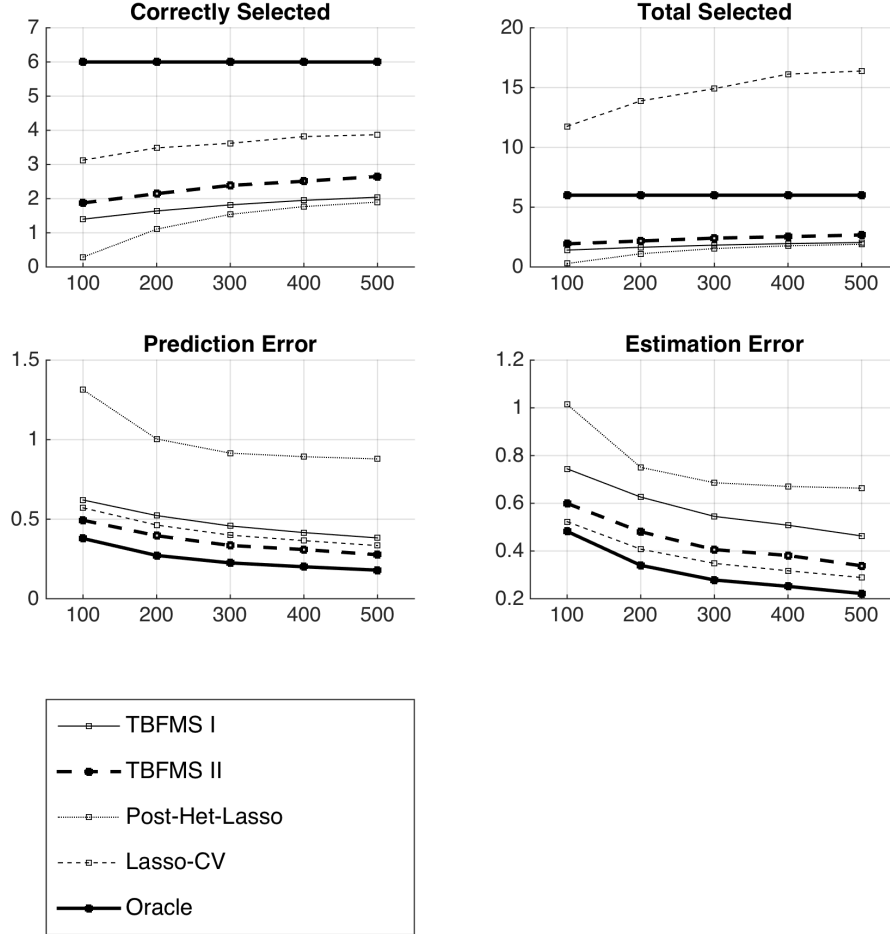


This figure presents simulation results for the estimators TBFMS I, TBFMS II, Post-Het-Lasso, Lasso-CV, and Oracle. The simulation design is described fully in Table 1. This figure presents estimates for the settings

- $s_0 = 6$: High Sparsity
- $b_0 = 0.5$: Positive Coefficients
- $\rho_0 = 0$: Homoskedastic disturbances.

Plots are based on 1000 simulation replications for every $n = 100, 200, 300, 400, 500$ indexed on the horizontal axis. The upper-left plot displays the number of correctly identified covariates from S_0 , averaged over simulation replications. The upper-right plot displays the total number of selected covariates, averaged over simulation replications. The bottom-left plot displays the prediction error defined as $\mathbb{E}_n[(x_i'\theta_0 - x_i'\hat{\theta})^2]^{1/2}$, averaged over simulation replications. The bottom-right plot displays estimation error defined as $\|\hat{\theta}_2 - \theta_0\|_2$, averaged over simulation replications.

FIG 2. *Simulation Results*



This figure presents simulation results for the estimators TBFMS I, TBFMS II, Post-Het-Lasso, Lasso-CV, and Oracle. The simulation design is described fully in Table 1. This figure presents estimates for the settings

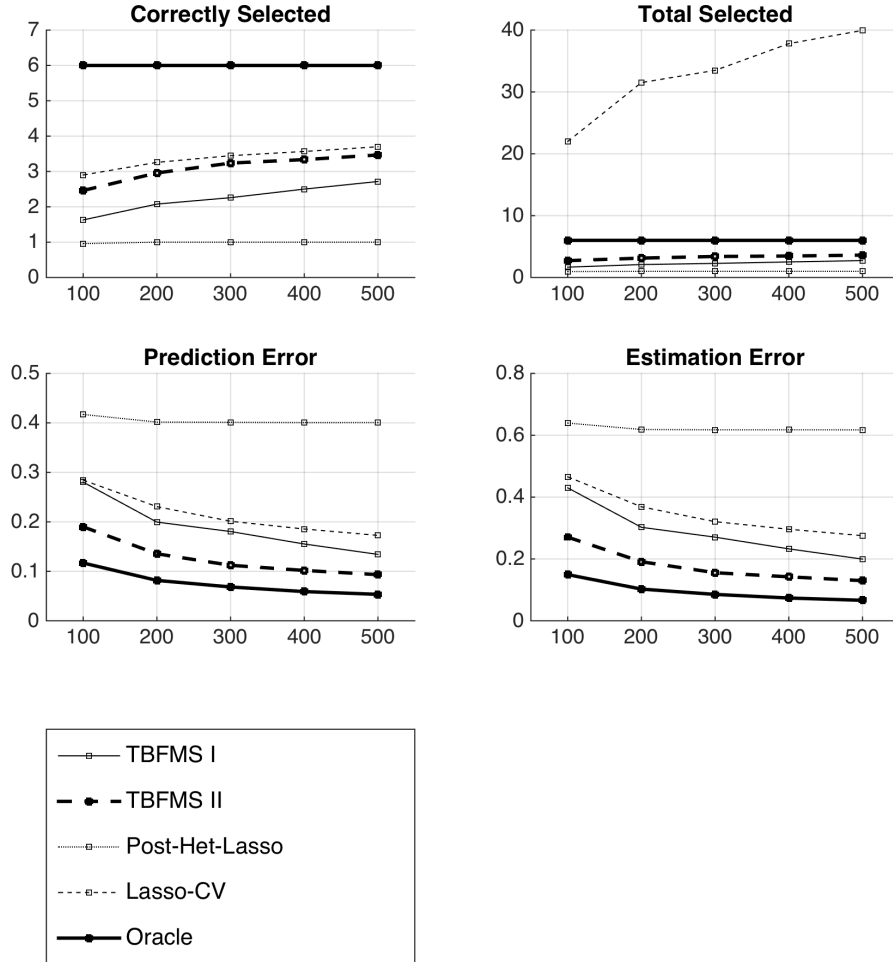
$s_0 = 6$: High Sparsity

$b_0 = 0.5$: Positive Coefficients

$\rho_0 = 0.5$: Heteroskedastic disturbances.

Plots are based on 1000 simulation replications for every $n = 100, 200, 300, 400, 500$ indexed on the horizontal axis. The upper-left plot displays the number of correctly identified covariates from S_0 , averaged over simulation replications. The upper-right plot displays the total number of selected covariates, averaged over simulation replications. The bottom-left plot displays the prediction error defined as $\mathbb{E}_n[(x'_i\theta_0 - x'_i\hat{\theta})^2]^{1/2}$, averaged over simulation replications. The bottom-right plot displays estimation error defined as $\|\hat{\theta}_2 - \theta_0\|_2$, averaged over simulation replications.

FIG 3. Simulation Results

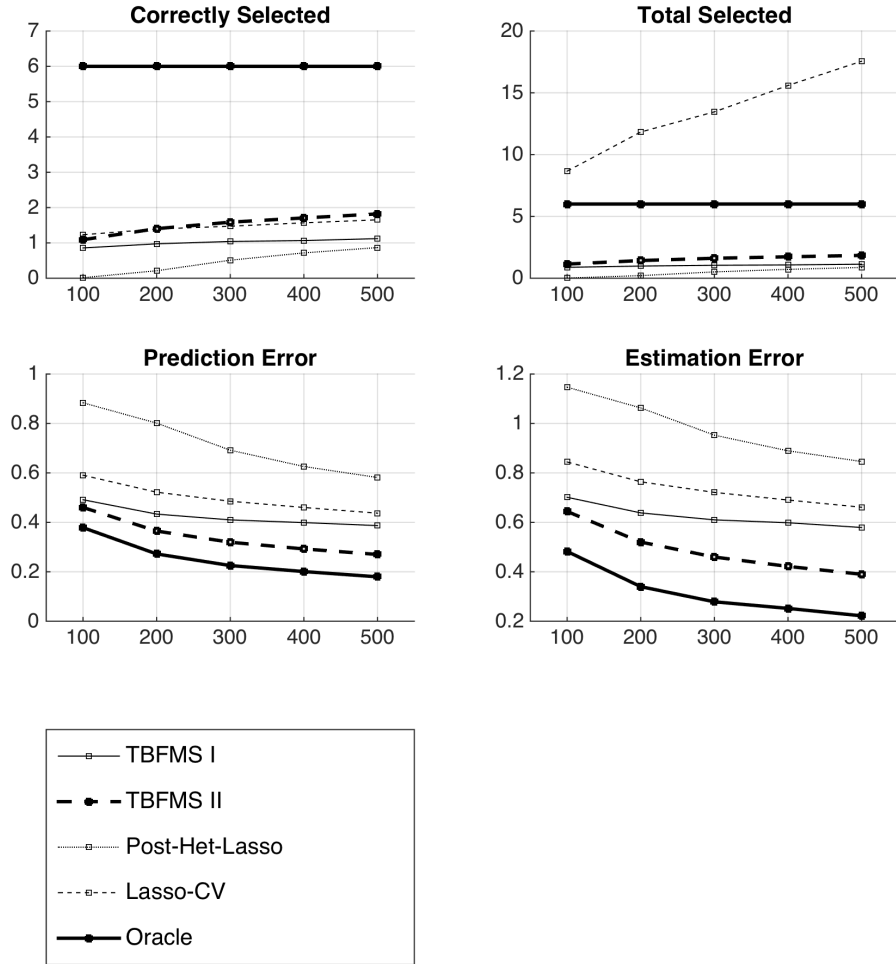


This figure presents simulation results for the estimators TBFMS I, TBFMS II, Post-Het-Lasso, Lasso-CV, and Oracle. The simulation design is described fully in Table 1. This figure presents estimates for the settings

- $s_0 = 6$: High Sparsity
- $b_0 = -0.5$: Alternating Coefficients
- $\rho_0 = 0$: Homoskedastic disturbances.

Plots are based on 1000 simulation replications for every $n = 100, 200, 300, 400, 500$ indexed on the horizontal axis. The upper-left plot displays the number of correctly identified covariates from S_0 , averaged over simulation replications. The upper-right plot displays the total number of selected covariates, averaged over simulation replications. The bottom-left plot displays the prediction error defined as $\mathbb{E}_n[(x_i'\theta_0 - x_i'\hat{\theta})^2]^{1/2}$, averaged over simulation replications. The bottom-right plot displays estimation error defined as $\|\hat{\theta}_2 - \theta_0\|_2$, averaged over simulation replications.

FIG 4. *Simulation Results*



This figure presents simulation results for the estimators TBFMS I, TBFMS II, Post-Het-Lasso, Lasso-CV, and Oracle. The simulation design is described fully in Table 1. This figure presents estimates for the settings

$s_0 = 6$: High Sparsity

$b_0 = -0.5$: Alternating Coefficients

$\rho_0 = 0.5$: Heteroskedastic disturbances.

Plots are based on 1000 simulation replications for every $n = 100, 200, 300, 400, 500$ indexed on the horizontal axis. The upper-left plot displays the number of correctly identified covariates from S_0 , averaged over simulation replications. The upper-right plot displays the total number of selected covariates, averaged over simulation replications. The bottom-left plot displays the prediction error defined as $\mathbb{E}_n[(x_i' \theta_0 - x_i' \hat{\theta})^2]^{1/2}$, averaged over simulation replications. The bottom-right plot displays estimation error defined as $\|\hat{\theta}_2 - \theta_0\|_2$, averaged over simulation replications.

4. Applications to structural inference in high-dimensional models

This section describes the application of TBFMS to inference for structural parameters in three commonly estimated economic problems.

- A.** The selection of instrumental variables for the estimation of the effect of an endogenous variable on an outcome of interest from a large set of potential instruments.
- B.** The selection of conditioning covariates for the estimation of the effect of a covariate of interest on an outcome from a large set of potential observable controls.
- C.** The selection of a conditioning set in the first stage of an instrumental variables regression from a large set of potential observable controls.

The general pattern in each of the above three models is that they contain a low dimensional parameter of interest and a high-dimensional nuisance parameter. In this sense they share many qualities with semiparametric estimation problems. Early works on the inference for a single coefficient in a high-dimensional regression model include [14], [71], [10], [9], [66], and [37], while early works for high-dimensional instrumental variables models include [6] and [30]. More detailed review of these two particular strands of literature are provided below. More general and more recent work in [25] develops theory for inference about a relatively low-dimensional set of prespecified target parameters when machine learning is used to estimate some features of the models specified by general moment conditions under weak conditions.

In all of the cases mentioned above, the quality of subsequent inference depends crucially on the quality of estimation of the various high-dimensional components that appear in the course of decomposing the problem into reduced form components. As a result, TBFMS is a good candidate for input into a larger structural estimation problem whenever it performs favorably in the relevant high-dimensional estimation steps in that problem.

The theorems that follow prove asymptotic normality for estimates of structural parameters of interest when instruments or conditioning variables are selected using TBFMS in an appropriate way. The formal results follow directly from the performance bounds for TBFMS given in Theorem 5 in conjunction with high-level conditions for inference in structural inference problems (see [14], [6], [25], [12] among others). Therefore, the main contribution of this section is in showing with simulation that TBFMS performs favorably relative to Lasso-based methods for some very simple data-generating processes. Subsections 4.1–4.2 briefly review estimation of Models **A** and **B** and provide theoretical justification for using TBFMS to estimate components of structural problems. Subsection 4.3 then conducts a simulation study to compare the finite sample performance of the use of

TBFMS and alternative estimation strategies in models **A** and **B**. Alternative methods include the two different implementations of Lasso considered in the previous section and an infeasible benchmark that knows the most important covariates in advance.

Finally, Subsection 4.5 presents an empirical application that fits model **C** to a classic data set of Acemoglu, Johnson, and Robinson.

4.1. Model A: Linear Instrumental Variables Model with High-Dimensional Instruments

Instrumental variables are a commonly used technique in applied econometrics. These methods give an important tool for estimating structural effects, but they are often imprecise. The precision of instrumental variables estimators can be improved by using many instruments or by trying to approximate the optimal instruments as in [2], [22], and [51].

This section follows the development [6], who consider using Post-Lasso to estimate optimal instruments. Using post-model selection methods to form first-stage predictions in IV estimation is a practical approach to obtaining the efficiency gains from using optimal instruments. The post-model selection approach simultaneously relieves the problems that arise with many instruments. An implication of Theorem 5 is that TBFMS provides good approximations to the optimal instruments when the total number of potential instruments is large.

Consider data given by $\mathcal{D}_n = \{(y_i, x_i, z_i)\}_{i=1}^n \sim P_n$ where $y_i \in \mathbb{R}$ are outcome variables, $x_i \in \mathbb{R}$ are endogenous variables of interest, and $z_i \in \mathbb{R}^p$ are instruments. Formally, the class **A** of instrumental variables models is defined to contain joint distribution P (for all n) such that

$$\begin{aligned} y_i &= \beta_0 x_i + \varepsilon_i \\ x_i &= z_i' \theta_0^A + u_i \end{aligned}$$

for some parameters $\beta_0 \in \mathbb{R}, \theta_0^A \in \mathbb{R}^p$ such that $E[\varepsilon_i | z_i] = E[u_i | z_i] = 0$ but possibly $E[\varepsilon_i u_i] \neq 0$.⁹

Consider estimation of the parameter of interest β_0 , the coefficient on the endogenous regressor, using TBMS to select instruments. Assume that the first-stage follows a sparse model with $|\text{support}(\theta_0^A)| \leq s_0 < n$. More precisely, consider the following condition.

⁹Extending this class of models to allow approximate sparsity is trivial by decomposing $u_i = u_i^a + u_i^o$, as was done in Conditions 4 and 5. Then $E[u_i^o | z_i] = 0$ can replace the requirement $E[u_i | z_i] = 0$.

The estimate $\widehat{\beta}$ of β_0 is defined as follows. Obtain $\widehat{\theta}^A$ from applying Algorithm 2 to $\{(x_i, z_i)\}_{i=1}^n$ using the hypothesis tests in Definition 1. Set $\widehat{x}_i = z_i' \widehat{\theta}^A$. Set

$$\widehat{\beta} = \widehat{Q}^{-1} \mathbb{E}_n[\widehat{x}_i y_i] \text{ where } \widehat{Q} = \mathbb{E}_n[x_i \widehat{x}_i]$$

$$\widehat{V} = \widehat{Q}^{-1} \widehat{\Omega} \widehat{Q}^{-1} \text{ where } \widehat{\Omega} = \mathbb{E}_n[\widehat{x}_i^2 \varepsilon_i^2].$$

The quantity \widehat{V} is used as an estimate of the variance of $\widehat{\beta}$. Under regularity conditions, it will be close to the quantity $V = Q^{-1} \Omega Q^{-1}$ with high probability where $Q = \mathbb{E}[(z_i \pi_0)^2]$, $\Omega = \mathbb{E}[(z_i \pi_0)^2 \varepsilon_i^2]$. The next theorem shows that the estimate $\widehat{\beta}$ is asymptotically Gaussian with variance V and \widehat{V} is consistent for V .

Condition 6 (*Regularity for Instrument Selection.*) $P_n \in \mathbf{A}$ satisfy Condition 6 for $\{(x_i, z_i)\}_{i=1}^n$ and Condition SM from [6].¹⁰

Theorem 6. *Consider data sets $\mathcal{D}_n = \{(y_i, x_i, z_i)\}_{i=1}^n \sim P_n$. Uniformly over all sequences $P_n \in \mathbf{A}$ for which Condition 6 holds, as $n \rightarrow \infty$,*

$$n^{1/2} V^{-1/2} (\widehat{\beta} - \beta_0) \rightarrow_d N(0, 1) \text{ and } V - \widehat{V} \rightarrow_p 0.$$

This theorem verifies that the IV estimator formed with instruments selected by TBFMS in a linear IV model is consistent and asymptotically normal. In addition, one can use the result with \widehat{V} defined above. Note that this inference will be valid uniformly over a large class of data-generating processes that include cases in which perfect instrument selection is impossible.

The conditions assumed in Condition 6 are fairly standard. The added conditions beyond the assertion that Condition 6 holds for $\{(x_i, z_i)\}_{i=1}^n$ are a simplified version of Condition SM from [6]. Outside of moment conditions, the main restriction is the assumption that the parameter β_0 would be strongly identified if $z_i \pi_0$ could be observed. Coupled with the sparse model, this condition implies that using a small number of the variables in z_i is

¹⁰For ease of reference, Condition SM from [6] is reproduced here, adapted to present notation. It states the following. $\mathbb{E}[\mathbb{E}_n[(z_i \theta_0^A)^2]]$ is bounded and away from zero, uniformly in n . The conditional variance $\mathbb{E}[\varepsilon_i^2 | z_i]$ is bounded a.s. uniformly from above and away from zero, uniformly in i and n . For some $q > 2$ and $q_\varepsilon > 0$, uniformly in n , $\max_{j \leq p} \mathbb{E}[\mathbb{E}_n[|z_{ij} \varepsilon_i|^3]] + \mathbb{E}[\mathbb{E}_n[|z_i' \theta_0^A|^q |\varepsilon_i|^{2q}]] + \mathbb{E}[\mathbb{E}_n[|z_i' \theta_0^A|^1]] + \mathbb{E}[\mathbb{E}_n[|\varepsilon_i|^{q_\varepsilon}]] + \mathbb{E}[|x_i|^q] = O(1)$. $\log(p) = o(n^{1/3})$. $s_0 \log(\max(p, n)) n^{2/q-1} = o(1)$. $s_0^2 \log^2(\max(p, n))/n = o(1)$. $\max_{j \leq p} \mathbb{E}_n[z_{ij}^2 \varepsilon_i^2] = O_P(1)$. Note that some of these conditions are redundant with Condition 5, but are nonetheless still stated for completeness.

sufficient to strongly identify β_0 , which rules out the case of weak-instruments as in [59] and many-weak-instruments as in [23].¹¹

4.2. Model B: Linear Model with High-Dimensional Control Variables

A different common strategy for identifying and estimating structural effects in economic research is based on assuming that covariates of interest are as good as randomly assigned conditional on observables. This leads to the practical problem researchers face of choosing which observed control covariates to include in the model. The high-dimensional setting provides a good formal framework for understanding data-dependent selection of control covariates. This section considers using TBFMS to select a set of covariates to include in a linear model from a large set of possible control variables.

The structure of the TBFMS procedure ensures that any coefficient that cannot be distinguished reliably from zero will be estimated to be exactly zero. Conversely, all estimated coefficients that are not set to zero can be statistically differentiated from zero, accounting for estimation noise. These properties complicate inference after model selection in sparse models that may have a set of variables with small but non-zero coefficients. In this case, use of TBFMS may result in excluding important conditioning covariates, which may lead to non-negligible omitted variables bias of parameters of interest. This intuition is formally developed in [54] and [44]. Offering solutions to this problem is the focus of a number of recent papers; see, for example, [8]; [6]; [71], [10], [9], [66]; [37], and [13].¹² In the previous references, Lasso and related shrinkage estimators were used in place of TBFMS. This section considers one technique for valid post-model selection inference, the Post-Double Selection technique ([14]), in conjunction with TBFMS.

Consider data given by $\mathcal{D}_n = \{(y_i, x_i, w_i)\}_{i=1}^n \sim P_n$ where $y_i \in \mathbb{R}$ are outcome variables, $x_i \in \mathbb{R}$ are variables of interest, and $w_i \in \mathbb{R}^p$ are controls. Formally, the class \mathbf{B} of linear models is defined to contain joint distribution P (for any n) such that

$$\begin{aligned} y_i &= x_i \beta_0 + w_i' \theta_0^{\mathbf{B}1} + \varepsilon_i \\ x_i &= w_i' \theta_0^{\mathbf{B}2} + u_i \end{aligned}$$

for some parameters $\beta_0 \in \mathbb{R}$, $\theta_0^{\mathbf{B}1}, \theta_0^{\mathbf{B}2} \in \mathbb{R}^p$ with $E[\varepsilon_i | w_i, x_i] = 0$ and $E[u_i | w_i] = 0$.¹³

Here, the impact of the policy/treatment variable x_i on the outcome y_i is measured by the parameter β_0 which is the target of inference. The variables w_i are potentially impor-

¹¹See also [34], who consider many-weak-instruments in a $p > n$ setting.

¹²These citations are ordered by date of first appearance on arXiv.

¹³As was discussed in the case with model \mathbf{A} , extending \mathbf{B} to allow approximate sparsity is trivial.

tant conditioning variables. Data are assumed independent but not necessarily identically distributed across i . The confounding factors w_i affect the policy variable via the function $w_i'\theta_0^{B_2}$ and the outcome variable via the function $w_i'\theta_0^{B_1}$. Both of the parameters $\theta_0^{B_1}$ and $\theta_0^{B_2}$ are unknown.

Inference about β_0 is impossible in this model without imposing further structure since $p > n$ elements in w_i are allowed. The additional structure is added by assuming that a sparsity condition applies to both $\theta_0^{B_1}$ and $\theta_0^{B_2}$. Once these assumptions are in place, the exogeneity of x_i may be taken as given after controlling linearly for a relatively small number, $s_0 < n$, of the variables in w_i whose identities are *a priori* unknown. Specifically, impose the following restrictions.

Condition 7 (*Regularity for Control Selection.*) $P_n \in \mathbf{B}$ satisfy Condition 5 for $\{(x_i, w_i)\}_{i=1}^n$ and $\{(y_i, w_i)\}_{i=1}^n$ and Condition SM from [14].¹⁴

To estimate β_0 in this environment, adopt the post-double-selection method of [9]. This method proceeds by first substituting to obtain predictive relationships for the outcome y_i and the treatment x_i in terms of only control variables:

$$\begin{aligned} y_i &= w_i'\theta_0^{\text{RF}} + v_i \\ x_i &= w_i'\theta_0^{\text{FS}} + u_i \end{aligned}$$

with $\theta_0^{\text{FS}} = \theta_0^{B_2}$ and $\theta_0^{\text{RF}} = \theta_0^{B_1} + \beta_0\theta_0^{B_2}$. Next use two variable selection steps. TBFMS is applied to each of the above two equations to select one set of variables that are useful for predicting y_i and another set of variables useful for predicting x_i . Once this is done, the union of the selected sets will index the final set of control variables.

[9] develop and discuss the post-double-selection method in detail. They note that including the union of the variables selected in each variable selection step helps address the issue that model selection is inherently prone to errors unless stringent assumptions are made. As noted by [44], the possibility of model selection mistakes precludes the possibility of valid post-model-selection inference based on a single Lasso regression within a large

¹⁴For ease of reference, Condition SM from [14] is reproduced here, adapted only to fit in with present notation. It states the following. For $(\check{y}_i, \check{\varepsilon}_i) = (y_i, v_i)$ and for $(\check{y}_i, \check{\varepsilon}_i) = (x_i, u_i)$ the following hold. $E[\mathbb{E}_n[|x_i|^q]] = O(1)$, $E[v_i^2|w_i, u_i]$ and $c E[u_i^2|w_i]$ are a.s. bounded by a constant above and away from zero, uniformly in i and n . $E[\mathbb{E}_n[|\check{\varepsilon}_i|^q]] + E[\mathbb{E}_n[\check{y}_i^2]] + \max_{j \leq p} E[\mathbb{E}_n[w_{ij}^2 \check{y}_i^2]] + E[\mathbb{E}_n[|w_{ij}^3 \check{\varepsilon}_i^3|]] + 1/E[\mathbb{E}_n[w_{ij}^2]] = O(1)$. $\log(p) = o(n^{1/3})$. $\max_{j \leq p} |(E_n - E\mathbb{E}_n)[w_{ij} 2\check{\varepsilon}_i^2]| + |(E_n - E\mathbb{E}_n)[w_{ij}^2 \check{y}_i^2]| + \max_{i \leq n} \|w_i\|_\infty^2 s_0 \log(\max(p, n))/n = o(1)$ with probability $1 - o(1)$. Note that some of these conditions are redundant with Condition 5, but are nonetheless still stated for completeness

class of interesting models. Using both model selection steps guards against such model selection mistakes and guarantees that the variables excluded in both model selection steps have a negligible contribution to omitted variables bias.

Formally, the estimate for $\hat{\beta}$ is given as follows. Obtain $\hat{\theta}^{\text{FS}}$ from applying Algorithm 2 to $\{(x_i, w_i)\}_{i=1}^n$ using the hypothesis tests in Definition 1. Set \hat{S}_{FS} . Obtain $\hat{\theta}^{\text{RF}}$ from applying Algorithm 2 to $\{(y_i, w_i)\}_{i=1}^n$ using the hypothesis tests in Definition 1. Set \hat{S}_{RF} . Set $\hat{S} = \hat{S}_{\text{FS}} \cup \hat{S}_{\text{RF}}$. Set

$$(\hat{\beta}, \hat{\theta}_{\hat{S}}^{\text{B1}}) = \hat{Q}^{-1} \mathbb{E}_n[(x_i, w_{i\hat{S}})y_i] \text{ where } \hat{Q} = \mathbb{E}_n[(x_i, w_{i\hat{S}})(x_i, w_{i\hat{S}})'].$$

$$\hat{V} = [\hat{Q}^{-1} \hat{\Omega} \hat{Q}^{-1}]_{11} \text{ where } \hat{\Omega} = \mathbb{E}_n[\hat{\varepsilon}_i^2(x_i, w_{i\hat{S}})(x_i, w_{i\hat{S}})']$$

and $\hat{\varepsilon}_i = y_i - \hat{\beta}x_i - w'_{i\hat{S}}\hat{\theta}_{\hat{S}}^{\text{B1}}$. Therefore, $\hat{\beta}$ is the OLS estimate obtained by regressing y_i on x_i and $w_{i\hat{S}}$, and \hat{V} is the conventional heteroskedasticity-robust variance estimate. As in the case of Model **A**, \hat{V} is an estimate of the variance of $\hat{\beta}$. Let $V = \mathbb{E}[u_i^2]^{-1} \mathbb{E}[u_i^2 v_i^2] \mathbb{E}[u_i^2]^{-1}$. The next theorem shows that $\hat{\beta}$ is asymptotically Gaussian and that $\hat{V} - V$ vanishes asymptotically in probability.

Theorem 7. *Consider sequences of data sets $\mathcal{D}_n = \{(y_i, x_i, w_i)\}_{i=1}^n \sim P_n$. Uniformly over all sequences $P_n \in \mathbf{B}$ for which Condition 7 holds, as $n \rightarrow \infty$,*

$$n^{1/2} V^{-1/2} (\hat{\beta} - \beta_0) \rightarrow_d \text{N}(0, 1) \text{ and } V - \hat{V} \rightarrow_p 0.$$

This theorem verifies that the OLS estimator that regresses y_i on x_i and the union of variables selected by TBFMS in the two stages described in Algorithm 4 is consistent and asymptotically normal with asymptotic variance that can be estimated with a heteroskedasticity-robust standard error estimator. Inference based on this result will be valid uniformly over a large class of data-generating processes, which includes cases in which perfect variable selection is impossible.

4.3. Post Model Selection Inference: Simulation Study

This section presents simulation evidence about the finite sample performance of inference about structural parameters. The simulation study provides a comparison with Lasso-based estimators as was done in the previous simulation study. The simulation design is divided into two parts, corresponding to Models **A** and **B** discussed above. The data-generating processes considered are given by Table 2.

TABLE 2
Structural Estimation Simulation Design

Structural Estimation Model A: High-Dimensional Instruments.

Data:	$\mathcal{D}_n = (y_i, x_i, z_i)_{i=1}^n \text{ iid}$
DGP:	$y_i = x_i \beta_0 + \varepsilon_i$ $x_i = z_i' \theta_0^A + u_i$ $p = \dim(z_i) = 2n$ $\theta_{0j}^A = b_0^{j-1} \mathbf{1}_{j \leq s_0}$ $z_{ij} \sim N(0, 1)$, with $\text{corr}(z_{ij}, z_{ik}) = 0.5^{ j-k }$ $(\varepsilon_i, u_i) \sim \sigma_i N\left(0, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$ $\sigma_i = \exp(\rho_0 \sum_{j=1}^p 0.75^{(p-j)} z_{ij})$
Settings:	$n \in \{100, 500\}$ $s_0 = 6$ $b_0 \in \{-0.5, 0.5\}$ $\rho_0 \in \{0, 1\}$
Target:	$\beta_0 = 1$

Structural Estimation Model B: High-Dimensional Controls.

Data:	$\mathcal{D}_n = (y_i, x_i, w_i)_{i=1}^n \text{ iid}$
DGP:	$y_i = x_i \beta_0 + w_i' \theta_0^{B1} + \varepsilon_i$ $x_i = w_i' \theta_0^{B2} + u_i$ $p = \dim(z_i) = 2n$ $\theta_{0j}^{B1} = b_0^{j-1} \mathbf{1}_{j \leq s_0}$ $\theta_{0j}^{B2} = \sin(j) \mathbf{1}_{j \leq s_0}$ $w_{ij} \sim N(0, 1)$, with $\text{corr}(w_{ij}, w_{ik}) = 0.5^{ j-k }$ $(\varepsilon_i, u_i) \sim \sigma_i N\left(0, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$ $\sigma_i = \exp(\rho_0 \sum_{j=1}^p 0.75^{(p-j)} z_{ij})$
Settings:	$n \in \{100, 500\}$ $s_0 = 6$ $b_0 \in \{-0.5, 0.5\}$ $\rho_0 \in \{0, 1\}$
Target:	$\beta_0 = 1$

Five estimators are considered for estimating both Models **A** and **B**, and are named TBFMS I, TBFMS II, Lasso-CV, Post-Het-Lasso, and Oracle, analogous to the simulation study of Section 3. The estimators differ only in that they replace TBFMS I with a different model selection technique in selecting covariates into the final estimated model as described in the previous sections. The respective model selection techniques used to replace TBFMS I in **A** and **B** are identical to those described in the simulation study of Section 3.

For each estimator and each simulation setting, the bias, the standard deviation of the point estimates, coverage probability and the average interval length are computed over the simulation replications. In the case of Model **A**, the number of times that no instruments were selected is also tracked. In these cases, no estimator of $\hat{\beta}$ is defined. The bias, standard deviation, coverage, and interval length calculations are made on only the replications for which at least one instrument was selected. Finally, final variance estimates \hat{V} for $\hat{\beta}$ are based on jackknife standard errors (more precisely, an approximation to jackknife standard errors given in [60] are used for model **A** and HC3 standard errors [69] are used for model **B**). Simulation results for Model **A** are shown in Table 3. Simulation results for Model **B** are shown in Table 4.

The simulation results indicate that across the data-generating processes considered, TBFMS I and TBFMS II generally achieve the good performance in terms of coverage probabilities. In addition, their bias, standard deviation, and interval lengths closely resemble the Oracle estimator. Note that in some simulations (most notably in Model **B**, Panel A) there is a large difference in coverage probabilities between the TBFMS I and II estimates and the Post-Het-Lasso estimate. Despite the fact that all the Post-Het-Lasso-based confidence sets are asymptotically uniformly valid, even against model selection mistakes, this example highlights the fact that finite sample model selection properties remain important considerations. In this example, the signal-to-noise ratio in the first stage is low, which is a setting in which detection of important variables is more difficult for Post-Het-Lasso relative to TBFMS I and II. Interestingly, in this case, using the relaxed penalty level with Lasso-CV does not help in terms of coverage probability.

TABLE 3
 Model A. Simulation Results: Instrument Selection

	$n = 100$				$n = 500$			
	Bias	StDev	Length	Cover	Bias	StDev	Length	Cover
A. $\rho_0 = 0$: Homoskedastic, $s_0 = 6$: High Sparsity, $b_0 = -0.5$: Alternating Sign								
TBFMS I	0.002	0.070	0.265	0.931	-0.002	0.030	0.111	0.934
TBFMS II	-0.004	0.067	0.252	0.925	-0.003	0.030	0.110	0.933
Post-Het-Lasso	0.003	0.073	0.281	0.944	-0.001	0.033	0.123	0.927
Lasso-CV	-0.035	0.066	0.240	0.873	-0.016	0.030	0.108	0.877
Oracle	-0.003	0.066	0.249	0.935	-0.002	0.030	0.110	0.933
B. $\rho_0 = 0.5$: Heteroskedastic, $s_0 = 6$: High Sparsity, $b_0 = -0.5$: Alternating Sign								
TBFMS I	-0.010	0.107	0.411	0.929	0.000	0.054	0.212	0.954
TBFMS II	-0.022	0.104	0.387	0.913	-0.006	0.050	0.195	0.940
Post-Het-Lasso	-0.057	0.069	0.264	0.750	-0.003	0.050	0.202	0.957
Lasso-CV	-0.150	0.146	0.387	0.586	-0.072	0.073	0.193	0.613
Oracle	-0.066	0.105	0.397	0.867	-0.015	0.049	0.187	0.929
C. $\rho_0 = 0$: Homoskedastic, $s_0 = 6$: High Sparsity, $b_0 = 0.5$: Positive Sign								
TBFMS I	-0.004	0.038	0.150	0.947	0.000	0.017	0.066	0.956
TBFMS II	-0.005	0.038	0.148	0.945	0.000	0.017	0.066	0.955
Post-Het-Lasso	-0.002	0.042	0.164	0.942	0.001	0.017	0.068	0.952
Lasso-CV	-0.010	0.037	0.146	0.934	-0.002	0.017	0.066	0.949
Oracle	-0.004	0.037	0.148	0.946	0.001	0.017	0.066	0.954
D. $\rho_0 = 0.5$: Heteroskedastic, $s_0 = 6$: High Sparsity, $b_0 = 0.5$: Positive Sign								
TBFMS I	-0.011	0.068	0.247	0.926	-0.002	0.030	0.116	0.941
TBFMS II	-0.014	0.067	0.240	0.916	-0.003	0.030	0.115	0.942
Post-Het-Lasso	-0.025	0.054	0.202	0.880	-0.000	0.030	0.118	0.954
Lasso-CV	-0.074	0.089	0.287	0.742	-0.022	0.035	0.114	0.845
Oracle	-0.026	0.065	0.240	0.913	-0.004	0.030	0.114	0.942

Simulation results for estimation in Model A. Simulations are conducted according to the design described in Table 2 with $s_0 = 6$, $b_0 \in \{-0.5, 0.5\}$, $\rho_0 \in \{0, 0.5\}$, and $n \in \{100, 500\}$. Estimates are presented for the five estimators, TBFMS I, TBFMS II, Post-Het-Lasso, Lasso-CV, and Oracle described in the text. The first column in each panel shows bias of the respective estimates for β_0 . The second column in each panel shows the standard deviation of the respective estimates for β_0 . The third column in each panel shows actual coverage probabilities of the respective 95% confidence intervals for β_0 . The fourth column in each panel shows length of confidence intervals for β_0 . The fifth column in each panel shows the total number of times across replications that no instruments were selected. Figures are based on 1000 simulation replications.

The quantities in this table are calculated over replications where at least 1 instrument was selected. TBFMS I and II each selected 0 instruments 5 times in Panel A, $n = 100$; 129 times in Panel B, $n = 100$; 7 times in Panel B, $n = 500$; 5 times in Panel D $n = 100$. Post-Het-Lasso selected 0 instruments 54 times in Panel A, $n = 100$; 992 times in Panel B, $n = 100$; 167 times in Panel B, $n = 500$; 708 times in Panel C, $n = 100$; 8 times in Panel C, $n = 500$; and 47 times in Panel D, $n = 100$. Lasso-CV selected 0 instruments 45 times in Panel B, $n = 100$; 1 time in Panel B, $n = 500$; 5 times in Panel D, $n = 100$.

TABLE 4
Model B Simulation Results: Control Selection in Linear Model

	$n = 100$				$n = 500$			
	Bias	StDev	Length	Cover	Bias	StDev	Length	Cover
A. $\rho_0 = 0$: Homoskedastic, $s_0 = 6$: High Sparsity, $b_0 = -0.5$: Alternating Sign								
TBFMS I	-0.052	0.114	0.377	0.845	-0.018	0.055	0.180	0.882
TBFMS II	0.006	0.108	0.429	0.949	0.002	0.045	0.179	0.957
Post-Het-Lasso	-0.190	0.056	0.211	0.067	-0.192	0.024	0.094	0.000
Lasso-CV	-0.193	0.054	0.215	0.069	-0.192	0.024	0.094	0.000
Oracle	0.006	0.099	0.419	0.958	0.001	0.044	0.178	0.955
B. $\rho_0 = 0.5$: Heteroskedastic, $s_0 = 6$: High Sparsity, $b_0 = -0.5$: Alternating Sign								
TBFMS I	0.015	0.324	1.044	0.888	-0.002	0.246	0.772	0.910
TBFMS II	-0.010	0.367	1.225	0.885	0.001	0.263	0.817	0.909
Post-Het-Lasso	-0.124	0.185	0.559	0.694	-0.105	0.156	0.430	0.630
Lasso-CV	-0.133	0.209	0.640	0.713	-0.106	0.157	0.430	0.630
Oracle	-0.014	0.391	1.340	0.886	0.005	0.265	0.830	0.911
C. $\rho_0 = 0$: Homoskedastic, $s_0 = 6$: High Sparsity, $b_0 = 0.5$: Positive Sign								
TBFMS I	-0.022	0.119	0.380	0.865	-0.038	0.061	0.182	0.785
TBFMS II	-0.003	0.118	0.430	0.913	0.000	0.045	0.179	0.954
Post-Het-Lasso	-0.085	0.061	0.233	0.682	-0.087	0.026	0.102	0.092
Lasso-CV	-0.083	0.060	0.233	0.690	-0.087	0.026	0.102	0.092
Oracle	-0.002	0.104	0.419	0.944	-0.001	0.045	0.178	0.956
D. $\rho_0 = 0.5$: Heteroskedastic, $s_0 = 6$: High Sparsity, $b_0 = 0.5$: Positive Sign								
TBFMS I	0.020	0.343	1.045	0.865	0.011	0.234	0.767	0.897
TBFMS II	0.026	0.376	1.226	0.877	0.003	0.248	0.817	0.905
Post-Het-Lasso	-0.138	0.216	0.595	0.668	-0.047	0.136	0.424	0.861
Lasso-CV	-0.037	0.202	0.633	0.904	-0.047	0.136	0.424	0.861
Oracle	0.025	0.396	1.341	0.898	0.007	0.251	0.829	0.908

Simulation results for estimation in Model **B**. Simulations are conducted according to the design described in Table 2 with $s_0 = 6$, $b_0 \in \{-0.5, 0.5\}$, $\rho_0 \in \{0, 0.5\}$, and $n \in \{100, 500\}$. Estimates are presented for the five estimators, TBFMS I, TBFMS II, Post-Het-Lasso, Lasso-CV, and Oracle described in the text. The first column in each panel shows bias of the respective estimates for β_0 . The second column in each panel shows standard deviation of the respective estimates for β_0 . The third column in each panel shows actual coverage probabilities of the respective 95% confidence intervals for β_0 . The fourth column in each panel shows length of confidence intervals for β_0 . Figures are based on 1000 simulation replications.

4.4. Empirical Illustration: Estimating the Effects of Institutions on Economic Output

In order to illustrate the use of TBFMS to help answer an empirical economic question, this section revisits 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 [12].

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 country's GDP. The primary candidate for such a factor discussed in [1] is geography. In this exercise, 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 principle, there are many ways to construct control variables related to a broad notion such as geography. These may include variables based on temperature, yearly rainfall, or terrain. In order to deal with the ambiguity of the definition of geography, construct a large set of different geographic variables. The strategy is to use TBFMS to choose from among the many variables and perform a subsequent IV analysis. Let w_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, latitude², latitude³, (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 [12]. Consider the model:

$$\log(\text{GDP per capita}_i) = \text{Protection from Expropriation}_i \theta_0 + w_i' \beta_0 + \varepsilon_i.$$

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. The data here uses the same set of 64 country-level observations as in [1]. When the set of control variables for geography, w_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 TBFMS gives a complementary analysis that chooses controls from among the constructed set of geographic variables.

Formally, cast the estimation over a model as belonging to the class \mathbf{C} which is characterized as follows. Consider data given by $\mathcal{D}_n = \{(y_i, x_i, z_i, w_i)\}_{i=1}^n \sim \mathbf{P}_n$ where $y_i \in \mathbb{R}$ is outcome variable, $x_i \in \mathbb{R}$ is an endogenous variable of interest, $z_i \in \mathbb{R}$ is an instrument, and $w_i \in \mathbb{R}^p$ are controls. Define the class \mathbf{C} of linear models to contain joint distribution \mathbf{P} (for all n) such that

$$\begin{aligned} y_i &= x_i \beta_0 + w_i' \theta_0^{\mathbf{C}1} + \varepsilon_i \\ x_i &= z_i' \pi_0 + w_i' \theta_0^{\mathbf{C}2} + u_i \\ z_i &= w_i' \theta_0^{\mathbf{C}3} + v_i \end{aligned}$$

for parameters $\beta_0, \pi_0 \in \mathbb{R}, \theta_0^{\mathbf{C}1}, \theta_0^{\mathbf{C}2}, \theta_0^{\mathbf{C}3} \in \mathbb{R}^p$ and $E[\varepsilon_i | z_i] = 0$.

Specializing to the current application, the fully expanded set of structural equations is given by the following three relations.

$$\log(\text{GDP per capita}_i) = \text{Protection from Expropriation}_i \theta_0 + w_i' \beta_0 + \varepsilon_i$$

$$\text{Protection from Expropriation}_i = \text{Settler Mortality}_i \pi_{01} + w_i' \Pi_{02} + v_i$$

$$\text{Settler Mortality}_i = w_i' \gamma_0 + u_i.$$

These yields three reduced form equations relating the structural variables to the controls.

$$\begin{aligned} \log(\text{GDP per capita}_i) &= w_i' \beta_0 + \tilde{\varepsilon}_i \\ \text{Protection from Expropriation}_i &= w_i' \tilde{\Pi}_{02} + \tilde{v}_i \\ \text{Settler Mortality}_i &= w_i' \gamma_0 + u_i. \end{aligned}$$

Select all geographic variables according to three steps. Let S_1 be the selected covariates from running TBFMS over the data $\log(\text{GDP per capita}_i)$ on w_i . Let S_2 be the selected covariates from running TBFMS over Protection from Expropriation $_i$ on, w_i . Let S_3 be the selected covariates from running TBFMS over Settler Mortality $_i$ on w_i . The final set of selected controls is given by $\hat{S} = S_1 \cup S_2 \cup S_3$. Valid estimation and inference of the structural parameter, θ , can then proceed by conventional IV estimation. As was the case in Models **A** and **B**, formal validity of this procedure can be justified uniformly over sequences of data sets generated from Model **C** (see [12] for example). The three model selection steps ensure that the final estimates are robust to classical concerns about pre-test biases.

Table 5 presents estimates. Each column presents first stage coefficients and final structural coefficients using a different method for estimating the model. Heteroskedasticity-robust standard errors are shown in paranthesis. 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 column 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 5, 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 5
Estimates: Effect of Institutions on Economic Growth

	Latitude	All Controls	TBFMS
First Stage	-0.537 (0.153)	-0.211 (0.217)	-0.380 (0.176)
Structural Estimate	0.969 (0.231)	0.984 (0.821)	0.835 (0.335)

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

*Note. The first two columns in this table (Latitude and All Controls) correspond to the first two columns in Table 5 of [12]. The exact quantities estimated in [12] are slightly different despite using identical data for the following reasons. Matlab and Stata differ in terms of coefficient estimates because of different regularization schemes and tolerances for nearly singular matrices (which also depend on the ordering of the variables). This difference is most relevant for the All Controls column. In order to facilitate replication, all of the quantities in this table have now been produced within a single statistical software, Matlab. The replication file is available from the author.

5. Conclusion

This paper has considered TBFMS for high-dimensional sparse linear regression problems. The procedure is shown to achieve estimation rates matching those of Lasso and Post-Lasso under a broad class of data-generating processes. In simulation studies, the method performs well in terms of prediction and as an input into larger structural inference problems.

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] Takeshi Amemiya. The non-linear two-stage least squares estimator. *Journal of Econometrics*, 2:105–110, 1974.
- [3] Donald W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, 1991.
- [4] J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146:304–317, 2008.
- [5] J. Bai and S. Ng. Boosting diffusion indices. *Journal of Applied Econometrics*, 24, 2009.
- [6] 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.
- [7] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. ArXiv, 2009.
- [8] A. Belloni, V. Chernozhukov, and C. Hansen. Lasso methods for gaussian instrumental variables models. 2010 arXiv:[math.ST], <http://arxiv.org/abs/1012.1297>, 2010.
- [9] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *ArXiv*, 2011. forthcoming, The Review of Economic Studies.
- [10] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III:245–295, 2013.
- [11] A. Belloni, V. Chernozhukov, C. Hansen, and D. Kozbur. Inference in high dimensional panel models with an application to gun control. *ArXiv:1411.6507*, 2014.
- [12] A. Belloni, C. Hansen, and V. Chernozhukov. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28:29–50, 2014.
- [13] Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. Program evaluation with high-dimensional data. *arXiv:1311.2645*, 2014.
- [14] 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.
- [15] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.
- [16] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.
- [17] Peter Bühlmann. Boosting for high-dimensional linear models. *Ann. Statist.*, 34(2):559–583, 2006.
- [18] F. Bunea, A. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
- [19] 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.
- [20] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.
- [21] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n . *Ann. Statist.*, 35(6):2313–2351, 2007.
- [22] G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34:305–334, 1987.
- [23] John C. Chao, Norman R. Swanson, Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen. Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments. *Econometric Theory*, 28(1):42–86, February 2012.
- [24] X. Chen, Q.-M. Shao, and W. Biao Wu. Self-normalized Cram\`er Type Moderate Deviations under Dependence. *ArXiv e-prints*, September 2014.
- [25] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. K. Newey, and J. Robins. Double Machine Learning for Treatment and Causal Parameters. *ArXiv e-prints*, July 2016.
- [26] D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated Lasso. *ArXiv e-prints*, May 2016.
- [27] 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.
- [28] W. Fithian, J. Taylor, R. Tibshirani, and R. Tibshirani. Selective Sequential Model Selection. *ArXiv e-prints*, December 2015.
- [29] Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics

- regression tools. *Technometrics*, 35(2):109–135, 1993.
- [30] E. Gautier and A. Tsybakov. High-dimensional instrumental variables regression and confidence sets. arXiv:1105.2454v2 [math.ST], 2011.
- [31] M. Grazier G’Sell, S. Wager, A. Chouldechova, and R. Tibshirani. Sequential Selection Procedures and False Discovery Rate Control. *ArXiv e-prints*, September 2013.
- [32] Alexander Grothendieck. Resume de la theorie metrique des produits tensoriels topologiques. *Soc. Mat. Sao-Paulo*, 8:1 – 79, 1953.
- [33] O. Guédon and R. Vershynin. Community detection in sparse networks via Grothendieck’s inequality. *ArXiv e-prints*, November 2014.
- [34] Christian Hansen and Damian Kozbur. Instrumental variables estimation with many weak instruments using regularized {JIVE}. *Journal of Econometrics*, 182(2):290 – 308, 2014.
- [35] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2009.
- [36] Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 2010.
- [37] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.
- [38] Bing-Yi Jing, Qi-Man Shao, and Qiyang Wang. Self-normalized cramer-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215, 2003.
- [39] Keith Knight. Shrinkage estimation for nearly singular designs. *Econometric Theory*, 24:323–337, 2008.
- [40] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.*, 45(1):7–57, 2009.
- [41] D. Kozbur. Inference in Additively Separable Models With a High Dimensional Set of Conditioning Variables. *ArXiv e-prints*, March 2015.
- [42] D. Kozbur. Sharp Convergence Rates for Forward Regression in High-Dimensional Sparse Linear Models. *ArXiv e-prints*, February 2017.
- [43] Damian Kozbur. Testing-based forward model selection. *American Economic Review*, 107(5):266–69, May 2017.
- [44] Hannes Leeb and Benedikt M. Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376, 2008.
- [45] A. Li and R. Foygel Barber. Accumulation tests for FDR control in ordered hypothesis testing. *ArXiv e-prints*, May 2015.

- [46] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.
- [47] 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.
- [48] Y. Luo and M. Spindler. *L2Boosting for Economic Applications*. *ArXiv e-prints*, February 2017.
- [49] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34:1436–1462, 2006.
- [50] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):2246–2270, 2009.
- [51] Whitney K. Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58:809–837, 1990.
- [52] Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.
- [53] G. Pisier. Grothendieck’s Theorem, past and present. *ArXiv e-prints*, January 2011.
- [54] Benedikt M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Sankhyā*, 71(1, Ser. A):1–18, 2009.
- [55] J. Qian, T. Hastie, R. Tibshirani, and N. Simon. *Glmnet for matlab (2013)*.
- [56] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *arXiv:0812.2818v1 [math.ST]*, 2008.
- [57] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *ArXiv:1106.1151*, 2011.
- [58] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:10251045, 2008.
- [59] D. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65:557–586, 1997.
- [60] Wurgler T. Steinhauer, A. Leverage and covariance matrix estimation in finite-sample iv regressions. *Working Paper 521, IEW, University of Zurich*, 2010.
- [61] T. Tao. When is correlation transitive? [<https://terrytao.wordpress.com/2014/06/05/when-is-correlation-transitive/>], June 2014.
- [62] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288, 1996.
- [63] R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact Post-Selection

- Inference for Sequential Regression Procedures. *ArXiv e-prints*, January 2014.
- [64] Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10):2231–2242, Oct 2004.
- [65] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.
- [66] 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.
- [67] 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.
- [68] Hansheng Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104:488:1512–1524, 2009.
- [69] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.
- [70] 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.
- [71] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.
- [72] Tong Zhang. On the consistency of feature selection using greedy least squares. *Journal of Machine Learning*, 10:555–568, 2009.
- [73] Tong Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Trans. Inf. Theor.*, 57(7):4689–4708, July 2011.

Appendix A: Proof of Theorem 1

Proof. The proof of Theorem 1 is divided into seven steps. Step 1 shows the first statement of Theorem 1. Step 2 defines a useful normalization of the selected covariates. Step 3 establishes certain bounds on the average correlation between selected covariates. Steps 4-6 show that if \widehat{s} is too high, then there must exist subsets of the selected covariates over which the average correlation must exceed what is permitted by assumption on the sparse eigenvalues of the empirical Gram matrix G . Step 7 concludes by pulling together the previous six steps.

Step 1

This first section of the proof provides a bound on $\mathbb{E}_n[(x'_i\theta_0 - x'_i\widehat{\theta})^2]$ which depends on \widehat{s} thereby proving the first statement of Theorem 1. First note that $\ell(\widehat{S}) = \ell(\widehat{S} \cup S_0) + [\ell(\widehat{S}) - \ell(\widehat{S} \cup S_0)]$. Note that $\ell(\widehat{S}) = \ell(\widehat{\theta})$ and $\ell(\widehat{S} \cup S_0) \leq \ell(\theta_0)$. In addition, by Lemma 3.3 of [27],

$$\ell(\widehat{S}) - \ell(\widehat{S} \cup S_0) \leq \varphi_{\min}(\widehat{s} + s_0)(G)^{-1} \sum_{j \in S_0 \setminus \widehat{S}} (-\Delta_j \ell(\widehat{S})) \leq s_0 t \varphi_{\min}(\widehat{s} + s_0)(G)^{-1}.$$

This gives

$$\ell(\widehat{\theta}) \leq \ell(\theta_0) + s_0 t \varphi_{\min}(\widehat{s} + s_0)(G)^{-1}.$$

Expanding the above two quadratics in $\ell(\cdot)$ gives

$$\begin{aligned} \mathbb{E}_n[(x'_i\theta_0 - x'_i\widehat{\theta})^2] &\leq |2\mathbb{E}_n[\varepsilon_i x'_i(\widehat{\theta} - \theta_0)]| + s_0 t \varphi_{\min}(\widehat{s} + s_0)(G)^{-1} \\ &\leq 2\|\mathbb{E}_n[\varepsilon_i x'_i]\|_{\infty} \|\theta_0 - \widehat{\theta}\|_1 + s_0 t \varphi_{\min}(\widehat{s} + s_0)(G)^{-1} \end{aligned}$$

To bound $\|\theta_0 - \widehat{\theta}\|_1$:

$$\begin{aligned} \|\theta_0 - \widehat{\theta}\|_1 &\leq \sqrt{\widehat{s} + s_0} \|\theta_0 - \widehat{\theta}\|_2 \\ &\leq \sqrt{\widehat{s} + s_0} \varphi_{\min}(\widehat{s} + s_0)(G)^{-1} \mathbb{E}_n[(x'_i\theta_0 - x'_i\widehat{\theta})^2]^{1/2}. \end{aligned}$$

Combining the above bounds and dividing by $\mathbb{E}_n[(x'_i\theta_0 - x'_i\widehat{\theta})^2]^{1/2}$ gives

$$\begin{aligned} \mathbb{E}_n[(x'_i\theta_0 - x'_i\widehat{\theta})^2]^{1/2} &\leq 2\|\mathbb{E}_n[\varepsilon_i x'_i]\|_{\infty} \sqrt{\widehat{s} + s_0} \varphi_{\min}(\widehat{s} + s_0)(G)^{-1} \\ &\quad + \frac{s_0 t \varphi_{\min}(\widehat{s} + s_0)(G)^{-1}}{\mathbb{E}_n[(x'_i\theta_0 - x'_i\widehat{\theta})^2]^{1/2}}. \end{aligned}$$

Finally, either $\mathbb{E}_n[(x'_i\theta_0 - x'_i\hat{\theta})^2]^{1/2} \leq \sqrt{s_0 t \varphi_{\min}(\hat{s} + s_0)(G)^{-1}}$, in which case the first statement of Theorem 1 holds, or alternatively $\mathbb{E}_n[(x'_i\theta_0 - x'_i\hat{\theta})^2]^{1/2} > \sqrt{s_0 t \varphi_{\min}(\hat{s} + s_0)(G)^{-1}}$, in which case

$$\begin{aligned} \mathbb{E}_n[(x'_i\theta - x'_i\hat{\theta})^2]^{1/2} &\leq 2\|\mathbb{E}_n[\varepsilon_i x'_i]\|_\infty \sqrt{\hat{s} + s_0} \varphi_{\min}(\hat{s} + s_0)(G)^{-1} \\ &\quad + \sqrt{s_0 t \varphi_{\min}(\hat{s} + s_0)(G)^{-1}} \end{aligned}$$

and the first statement of Theorem 1 follows.

Step 2

This section of the proof defines *true* and *false* covariates, introduces a convenient orthogonalization of all selected covariates, and associates to each false selected covariate a parameter $\tilde{\gamma}_j$ on which the analysis is based.

Let $x_j = [x_{1j}, \dots, x_{nj}]'$ be the vector in \mathbb{R}^n with components x_{ij} stacked vertically. Similarly, define $\varepsilon = [\varepsilon_1, \dots, \varepsilon_n]'$ and $y = [y_1, \dots, y_n]'$. Let $v_k \in \mathbb{R}^n$, $k = 1, \dots, s_0$ denote *true covariates* which are defined as the vectors x_j for $j \in S_0$. Define *false covariates* simply as those which do not belong to S_0 .

Consider any point in time in the the Simple Forward Selection algorithm when there are m false covariates selected into the model. These falsely selected covariates are denoted w_1, \dots, w_m , each in \mathbb{R}^n , ordered according to the order they were selected.

The true covariates are also ordered according to the order they are selected into the model. Any true covariates unselected after the m false covariate selection are temporarily ordered arbitrarily at the end of the list. Let \mathcal{M}_k be projection in \mathbb{R}^n onto the space orthogonal to $\text{span}(\{v_1, \dots, v_k\})$. Let

$$\tilde{v}_k = \frac{\mathcal{M}_{k-1}v_k}{(v'_k \mathcal{M}_{k-1}v_k)^{1/2}} \text{ for } k = 1, \dots, s_0.$$

In addition, set

$$\tilde{\varepsilon} = \frac{\mathcal{M}_{s_0}\varepsilon}{(\varepsilon' \mathcal{M}_{s_0}\varepsilon)^{1/2}}.$$

Let $\tilde{V}_{\text{temp}} = [\tilde{v}_1, \dots, \tilde{v}_{s_0}]$, ordered according to the temporary order. Note that there is $\tilde{\theta} \in \mathbb{R}^{s_0}$ and $\tilde{\theta}_\varepsilon \in \mathbb{R}$ such that

$$\tilde{V}_{\text{temp}}\tilde{\theta}_{\text{temp}} + \tilde{\theta}_\varepsilon\tilde{\varepsilon} = y.$$

At this time, reorder the true covariates. Let \hat{k} denote the index of the final true covariate selected into the model when the m -th false covariate is selected. The variables $\tilde{v}_1, \dots, \tilde{v}_{\hat{k}}$ maintain their original order. The unselected true covariates $\tilde{v}_{\hat{k}+1}, \dots, \tilde{v}_{s_0}$ are reordered in

such a way that under the new ordering, $\tilde{\theta}_{k,\text{temp}} \geq \tilde{\theta}_{l,\text{temp}}$ whenever $l > k$. Also define $\tilde{V} = [\tilde{v}_1, \dots, \tilde{v}_{s_0}]$ consistent with the new ordering. Redefine $\tilde{\theta}$ by $\tilde{V}\tilde{\theta} + \tilde{\theta}_\varepsilon\tilde{\varepsilon} = y$ so that it is also consistent with the new ordering. Note that no new orthogonalization needs to be done.

For any set S , Let \mathcal{Q}_S be projection onto the space orthogonal to $\text{span}(\{x_j, j \in S\})$. For each selected covariate, w_j , set $S_{\text{pre-}w_j}$ to be the set of (both true and false) covariates selected prior to w_j . Define

$$\tilde{w}_j = c_j \mathcal{Q}_{S_{\text{pre-}w_j}} w_j$$

where the normalization constants c_j are defined in the next paragraph.

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 $\tilde{u}_j \in \text{span}(\tilde{V})^\perp$. The normalizations c_j introduced above are then chosen so that $\tilde{u}_j' \tilde{u}_j = 1$.

Associates to each false covariate \tilde{w}_j , a vector $\tilde{\gamma}_j \in \mathbb{R}^{s_0}$, defined as the solution in \mathbb{R}^{s_0} to the following equation

$$\tilde{V}\tilde{\gamma}_j = \tilde{r}_j.$$

Set $\tilde{\gamma}_{j\varepsilon} = \tilde{\varepsilon}' \tilde{w}_j$. Assume without loss of generality that each component of $\tilde{\theta}$ is positive (since otherwise, the true covariates can just be multiplied by -1 .) Also assume without loss of generality that $\tilde{\gamma}_j' \tilde{\theta} \geq 0$.

Step 3

This section provides upper bounds on quantities related to the $\tilde{\gamma}_j$ defined above. The idea guiding the argument in the next sections is that if too many covariates w_j are selected, then on average they must be correlated with each other since they must be correlated to y . For a discussion of partial transitivity of correlation, see [61]. If the covariates are highly correlated amongst themselves, then $\varphi_{\min}(m + s_0)(G)^{-1}$ must be very high. As a result, the sparse eigenvalues of G can be used to upper bound the number of selections. Average correlations between covariates are tracked with the aid of the quantities $\tilde{\gamma}_j$.

Divide the set of false covariates into two sets A_1 and A_2 where

$$A_1 = \left\{ j : |\tilde{\gamma}_{j\varepsilon}| \leq \frac{t^{1/2}n^{1/2}}{(2\varepsilon'\mathcal{M}_{s_0}\varepsilon)^{1/2}} \right\}, \quad A_2 = \left\{ j : |\tilde{\gamma}_{j\varepsilon}| > \frac{t^{1/2}n^{1/2}}{(2\varepsilon'\mathcal{M}_{s_0}\varepsilon)^{1/2}} \right\}.$$

Sections 3 - 5 of the proof bound the number of elements in A_1 . Section 6 of the proof bounds the number of elements in A_2 .

Suppose the set A_1 contains m_1 total false selections. Collect these false selections into $\tilde{W} = [\tilde{w}_{j_1}, \dots, \tilde{w}_{j_{m_1}}]$. Set $\tilde{R} = [\tilde{r}_{j_1}, \dots, \tilde{r}_{j_{m_1}}]$, $\tilde{U} = [\tilde{u}_{j_1}, \dots, \tilde{u}_{j_{m_1}}]$. Decompose $\tilde{W} = \tilde{R} + \tilde{U}$.

Then $\tilde{W}'\tilde{W} = \tilde{R}'\tilde{R} + \tilde{U}'\tilde{U}$. Since $\text{diag}(\tilde{U}'\tilde{U}) = I$, it follows that the average inner product between the \tilde{u}_j , given by $\bar{\rho}$:

$$\bar{\rho} = \frac{1}{m_1(m_1 - 1)} \sum_{j \neq l \in A_1} \tilde{u}_j' \tilde{u}_l,$$

must be bounded below by

$$\bar{\rho} \geq -\frac{1}{m_1 - 1}$$

due to the positive definiteness of $\tilde{U}'\tilde{U}$. This implies an upper bound on the average off-diagonal term in $\tilde{R}'\tilde{R}$ since $\tilde{W}'\tilde{W}$ is a diagonal matrix. Since \tilde{v}_k are orthonormal, the sum of all the elements of $\tilde{R}'\tilde{R}$ is given by $\|\sum_{j \in A_1} \tilde{\gamma}_j\|_2^2$. Since $\|\sum_{j \in A_1} \tilde{\gamma}_j\|_2^2 = \sum_{j \in A_1} \|\tilde{\gamma}_j'\|_2^2 + \sum_{j \neq l \in A_1} \tilde{\gamma}_j' \tilde{\gamma}_l$ and since $\tilde{W}'\tilde{W}$ is a diagonal matrix, it must be the case that

$$\frac{1}{m_1(m_1 - 1)} \sum_{j \neq l \in A_1} \tilde{\gamma}_j' \tilde{\gamma}_l = -\bar{\rho}.$$

Therefore,

$$\bar{\rho} = \frac{1}{m_1(m_1 - 1)} \left(\left\| \sum_{j \in A_1} \tilde{\gamma}_j \right\|_2^2 - \sum_{j \in A_1} \|\tilde{\gamma}_j\|_2^2 \right) \leq \frac{1}{m_1 - 1}.$$

This implies that

$$\left\| \sum_{j \in A_1} \tilde{\gamma}_j \right\|_2^2 \leq m_1 + \sum_{j \in A_1} \|\tilde{\gamma}_j\|_2^2.$$

Next, bound $\max_{j \in A_1} \|\tilde{\gamma}_j\|_2^2$. Note $\|\tilde{\gamma}_j\|_2^2 = \|\tilde{r}_j\|_2^2$ since \tilde{V} is orthonormal. Note that $\|\tilde{u}_j\|_2^2 / \|\tilde{w}_j\|_2^2 = 1 / \|\tilde{w}_j\|_2^2$ is lower bounded by $\varphi_{\min}(m + s_0)(G)$. This follows from the fact that you can associate $\|\tilde{u}_j/c_j\|_2^2$ to an element of a the inverse covariance matrix for w_j and previously selected covariates. Therefore, $\|\tilde{r}_j\|_2^2 = \|\tilde{w}_j\|_2^2 - 1 \leq \varphi_{\min}(m + s_0)(G)^{-1} - 1$. It follows that

$$\max_{j \in A_1} \|\tilde{\gamma}_j\|_2^2 \leq \varphi_{\min}(m + s_0)(G)^{-1} - 1.$$

This then implies that

$$\left\| \sum_{j \in A_1} \tilde{\gamma}_j \right\|_2^2 \leq m_1 \varphi_{\min}(m + s_0)(G)^{-1}.$$

The same argument as above also shows that for any choice $e_j \in \{-1, 1\}$ of signs, it is always the case that

$$\left\| \sum_{j \in A_1} e_j \tilde{\gamma}_j \right\|_2^2 \leq m_1 \varphi_{\min}(m + s_0)(G)^{-1}.$$

(In more detail, take $\tilde{W}_e = [\tilde{w}_{j_1} e_{j_1}, \dots, \tilde{w}_{j_{m_1}} e_{j_{m_1}}]$, etc. and rerun the same argument.)

Step 4

Next search for a particular choice of signs $\{e_j\}_{j \in A_1}$ which give a lower bound proportional to m_1^2/s_0 on the above term. Note that this will imply an upper bound on m_1 . For each $k = 1, \dots, s_0$, let A_{1k} be the set which contains those $j \in A_1$ such that w_j is selected before v_k , but not before any other true covariate. Note that the sets $A_{1(\widehat{k}+2)}, \dots, A_{1(s_0+1)}$ are set empty if $\widehat{k} < s_0$. Also, empty sums are set to zero. Define the following two matrices:

$$\Gamma = \begin{bmatrix} \sum_{j \in A_{11}} \tilde{\gamma}_{j1} & \sum_{j \in A_{11}} \tilde{\gamma}_{j2} & \dots & \sum_{j \in A_{11}} \tilde{\gamma}_{js_0} \\ 0 & \sum_{j \in A_{12}} \tilde{\gamma}_{j2} & \dots & \sum_{j \in A_{12}} \tilde{\gamma}_{js_0} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sum_{j \in A_{1s_0}} \tilde{\gamma}_{js_0} \end{bmatrix}, \quad B = \begin{bmatrix} \frac{\tilde{\theta}_1}{\tilde{\theta}_1} & \frac{\tilde{\theta}_2}{\tilde{\theta}_1} & \dots & \frac{\tilde{\theta}_{s_0}}{\tilde{\theta}_1} \\ \frac{\tilde{\theta}_2}{\tilde{\theta}_1} & \frac{\tilde{\theta}_2}{\tilde{\theta}_2} & \dots & \frac{\tilde{\theta}_{s_0}}{\tilde{\theta}_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\tilde{\theta}_{s_0}}{\tilde{\theta}_1} & \frac{\tilde{\theta}_{s_0}}{\tilde{\theta}_2} & \dots & \frac{\tilde{\theta}_{s_0}}{\tilde{\theta}_{s_0}} \end{bmatrix}$$

Note that the k th row of Γ is equal to $\sum_{j \in A_{1k}} \tilde{\gamma}_k$ since the orthogonalization process had enforced $\tilde{\gamma}_{jl} = 0$ for each $l < k$. Therefore, the diagonal elements of the product ΓB satisfy the equality

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

Let C_1, C_2 be constants such that

$$\tilde{\gamma}'_j \tilde{\theta} / \tilde{\theta}_k \geq C_1$$

for $j \in A_{1k}$, and

$$\tilde{\theta}_k / \tilde{\theta}_l \geq C_2$$

for $l > k$. These key constants are calculated explicitly in Section 5 of the proof. They imply that

$$[\Gamma B]_{k,k} \geq C_1 |A_{1k}| \quad \text{and} \quad \text{tr}(\Gamma B) \geq C_1 m_1.$$

Further observe that whenever $\tilde{\theta}_k \geq C_2 \tilde{\theta}_l$ for each $k, l > k$, assuming without loss of generality that $C_2 \leq 1$, that $(B + C_2^{-1}I)$ is positive semidefinite. 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 $\tilde{\theta}_k / \tilde{\theta}_{k-1}$ covariance with the k th random variable. Then $B + C_2^{-1}I$ has a positive definite symmetric matrix square

root so let $D^2 = B + C_2^{-1}I$. 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 $\leq 1 + C_2^{-1}$ and therefore B decomposes into a product $B = E'F$ where the rows of E, F have norms bounded by $1 + C_2^{-1} + C_2^{-1/2}$. Therefore, let $C_3 = 1 + C_2^{-1} + C_2^{-1/2}$.

Consider the set

$$\mathcal{G}_{s_0} = \{Z \in \mathbb{R}^{s_0 \times s_0} : Z_{ij} = X_i'Y_j \text{ for some } X_i, Y_j \in \mathbb{R}^{s_0}, \|X_i\|_2, \|Y_j\|_2 \leq 1\}$$

and observe that $\bar{B} := C_3^{-1}B \in \mathcal{G}_{s_0}$. Then this observation allows the use of Grothendieck's inequality (using the exact form described in [33]) which gives

$$\max_{Z \in \mathcal{G}_{s_0}} \text{tr}(\Gamma Z) \leq K_G^{\mathbb{R}} \|\Gamma'\|_{\infty \rightarrow 1}.$$

Here, $K_G^{\mathbb{R}}$ is an absolute constant which is known to be less than 1.783. It does not depend on s_0 . Therefore, $C_1 m \leq \text{tr}(\Gamma B) = C_3 \text{tr}(\Gamma \bar{B}) \leq \max_{Z \in \mathcal{G}_{s_0}} \text{tr}(\Gamma Z) \leq K_G^{\mathbb{R}} \|\Gamma'\|_{\infty \rightarrow 1}$, which implies

$$\left(K_G^{\mathbb{R}}\right)^{-1} C_3^{-1} C_1 m_1 \leq \|\Gamma'\|_{\infty \rightarrow 1}.$$

Therefore, there is $\nu \in \{-1, 1\}^{s_0}$ such that $\|\nu'\Gamma\|_1 \geq \left(K_G^{\mathbb{R}}\right)^{-1} C_3^{-1} C_1 m_1$. For this particular choice of ν , it follows that

$$\|\nu'\Gamma\|_2 \geq s_0^{-1/2} \left(K_G^{\mathbb{R}}\right)^{-1} C_3^{-1} C_1 m_1.$$

Then by definition of Γ , $\|\nu'\Gamma\|_2^2 = \|\sum_{k=1}^{s_0} \sum_{j \in A_{1k}} \nu_k \tilde{\gamma}_j\|_2^2$. In Section 3, it was noted that $\|\sum_{j=1}^{m_1} e_j \tilde{\gamma}_j\|_2^2 \leq m_1 \varphi_{\min}(m + s_0)(G)^{-1}$ for any choice of signs $e_j \in \{-1, 1\}^{m_1}$. It follows that

$$s_0^{-1} \left(K_G^{\mathbb{R}}\right)^{-2} C_3^{-2} C_1^2 m_1^2 \leq m_1 \varphi_{\min}(m + s_0)(G)^{-1}$$

which yields the conclusion

$$m_1 \leq \varphi_{\min}(m + s_0)(G)^{-1} C_1^{-2} C_3^2 \left(K_G^{\mathbb{R}}\right)^2 s_0.$$

Step 5

It is left to calculate C_1, C_2 which lower bound $\tilde{\gamma}_j'/\tilde{\theta}_k$ for $j \in A_{1k}$ and $\tilde{\theta}_k/\tilde{\theta}_l$ for $l > k$. A simple derivation can be made to show that the incremental decrease in empirical loss from the j th false selection is

$$-\Delta_j \ell(S_{\text{pre-}w_j}) = \frac{1}{n} y' \tilde{w}_j (\tilde{w}_j' \tilde{w}_j)^{-1} \tilde{w}_j' y = \frac{1}{n} \frac{1}{\tilde{w}_j' \tilde{w}_j} (\tilde{\theta}' \tilde{\gamma}_j + \tilde{\theta}'_{\varepsilon} \tilde{\gamma}_{j\varepsilon})^2$$

Note the slight abuse of notation in $-\Delta_j(S_{\text{pre-}w_j})$ signifying change in loss under inclusion of w_j rather than x_j . Next,

$$(\tilde{\theta}'\tilde{\gamma}_j + \tilde{\theta}'_{\tilde{\varepsilon}}\tilde{\gamma}_{j\tilde{\varepsilon}})^2 \leq 2(\tilde{\theta}'\tilde{\gamma}_j)^2 + 2(\tilde{\theta}'_{\tilde{\varepsilon}}\tilde{\gamma}_{j\tilde{\varepsilon}})^2$$

Since $\tilde{\theta}_{\tilde{\varepsilon}} = (\varepsilon'\mathcal{M}_{s_0\varepsilon})^{1/2}$, $\tilde{w}'_j\tilde{w}_j \geq 1$, and $j \in A_1$ it follows that

$$\frac{1}{n} \frac{1}{\tilde{w}'_j\tilde{w}_j} (\tilde{\theta}'_{\tilde{\varepsilon}}\tilde{\gamma}_{j\tilde{\varepsilon}})^2 \leq \frac{1}{n} \frac{1}{\tilde{w}'_j\tilde{w}_j} \tilde{\theta}_{\tilde{\varepsilon}}^2 \left(\frac{t^{1/2}n^{1/2}}{2(\varepsilon'\mathcal{M}_{s_0\varepsilon})^{1/2}} \right)^2 \leq \frac{t}{4}.$$

This implies

$$\frac{1}{2}(-\Delta_j\ell(S_{\text{pre-}w_j})) \leq \frac{1}{n} \frac{1}{\tilde{w}'_j\tilde{w}_j} (\tilde{\theta}'\tilde{\gamma}_j)^2 + \frac{t}{4}.$$

By the condition that the false j is selected, it holds that $-\Delta_j\ell(S_{\text{pre-}w_j}) > t$ and so $\frac{1}{4}(-\Delta_j\ell(S_{\text{pre-}w_j})) > \frac{t}{4}$ which implies that

$$\frac{1}{2}(-\Delta_j\ell(S_{\text{pre-}w_j})) - \frac{t}{4} \geq \frac{1}{4}(-\Delta_j\ell(S_{\text{pre-}w_j})).$$

Finally, this yields that

$$\frac{1}{n\tilde{w}'_j\tilde{w}_j} (\tilde{\gamma}'_j\tilde{\theta})^2 \geq \frac{1}{4}(-\Delta_j\ell(S_{\text{pre-}w_j})).$$

By the fact that w_j was selected ahead of v_k it holds that

$$-\Delta_j\ell(S_{\text{pre-}w_j}) \geq -\Delta_k\ell(S_{\text{pre-}w_j}).$$

Therefore, further bound the righthand side. Let \tilde{z}_k be the projection of \tilde{v}_k onto the space orthogonal to all previously selected (true and false) covariates. Then

$$-\Delta_k\ell(S_{\text{pre-}w_j}) \geq \frac{1}{n} \tilde{z}'_k \tilde{z}_k \tilde{\theta}_k^2.$$

Furthermore, $\tilde{z}'_k \tilde{z}_k \geq \varphi_{\min}(m + s_0)(G)^2$. This is seen by noting that \tilde{z}_k results in the composition of two projections onto a span of covariates of size bounded by $m + s_0$.

This gives

$$\frac{1}{n\tilde{w}'_j\tilde{w}_j} (\tilde{\gamma}'_j\tilde{\theta})^2 \geq \frac{1}{4} \frac{1}{n} \varphi_{\min}(m + s_0)(G)^2 \tilde{\theta}_k^2.$$

Using the fact that $\tilde{w}'_j\tilde{w}_j \geq 1$ implies that

$$(\tilde{\gamma}'_j\tilde{\theta})^2 / \tilde{\theta}_k^2 \geq \frac{1}{4} \varphi_{\min}(m + s_0)(G)^2.$$

Now suppose no true variables remain when j is selected, then $\tilde{w}'_j\tilde{w}_j = \tilde{u}'_j\tilde{u}_j = 1$. Therefore,

$$-\Delta_j \ell(S_{\text{pre-}w_j}) = \frac{1}{n} \tilde{\gamma}_{j\tilde{\varepsilon}}^2 \tilde{\theta}_{\tilde{\varepsilon}}^2 \geq t$$

Note that $\tilde{\theta}$ is given by $\tilde{\theta}_{\tilde{\varepsilon}} = \tilde{\varepsilon}' y = \varepsilon' \mathcal{M}_{s_0} y / (\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2} = (\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}$. Therefore,

$$\tilde{\gamma}_{j\tilde{\varepsilon}}^2 \geq t \frac{n}{\varepsilon' \mathcal{M}_{s_0} \varepsilon}.$$

This implies that $j \in A_2$. Therefore, set $C_1 = \frac{1}{2} \varphi_{\min}(m + s_0)(G)$.

Next, construct C_2 . For each selected true covariate, v_k , set $S_{\text{pre-}v_k}$ to be the set of (both true and false) covariates selected prior to v_k . Note that

$$\tilde{\theta}_k^2 = -\Delta_k \ell(\{v_1, \dots, v_{k-1}\}) \geq -\Delta_k \ell(S_{\text{pre-}v_k})$$

since $\{v_1, \dots, v_{k-1}\} \subseteq S_{\text{pre-}v_k}$. In addition, if v_k is selected before v_l (or v_l is not selected), then

$$-\Delta_k \ell(S_{\text{pre-}v_k}) \geq -\Delta_l \ell(S_{\text{pre-}v_k}) \geq \tilde{z}_l' \tilde{z}_l \tilde{\theta}_l^2 \geq \varphi_{\min}(\hat{s} + s_0)(G)^2 \tilde{\theta}_l^2.$$

Therefore, taking

$$C_2 = \varphi_{\min}(m + s_0)(G)$$

implies that $\tilde{\theta}_k / \tilde{\theta}_l \geq C_2$ for any $l > k$.

Step 6

In this section, the number of elements of A_2 is bounded. Recall that the criteria for $j \in A_2$ is that $|\tilde{\gamma}_{j\tilde{\varepsilon}}| > \frac{t^{1/2} n^{1/2}}{(2\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}}$. Note also that $\tilde{\gamma}_{j\tilde{\varepsilon}}$ is found by the coefficient in the expression

$$\tilde{\gamma}_{j\tilde{\varepsilon}} = \tilde{\varepsilon}' \tilde{w}_j = \varepsilon' \frac{1}{(\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}} \mathcal{M}_{s_0} \tilde{w}_j$$

Next, let H be the matrix $H = [v_1, \dots, v_{s_0}, w_1, \dots, w_m]$. Note that

$$\frac{1}{(\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}} \mathcal{M}_{s_0} \tilde{w}_j \in \text{span}(H)$$

Which implies that the above expression is unchanged when premultiplied by $H(H'H)^{-1}H'$. Therefore,

$$\tilde{\gamma}_{j\tilde{\varepsilon}} = \varepsilon' H(H'H)^{-1}H' \frac{1}{(\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}} \mathcal{M}_{s_0} \tilde{w}_j.$$

Let μ_j be the +1 for each $j \in A_2$ such that $\tilde{\gamma}_{j\tilde{\varepsilon}} > 0$ and -1 for each $j \in A_2$ such that $\tilde{\gamma}_{j\tilde{\varepsilon}} < 0$. By the fact that $j \in A_2$, $\tilde{\gamma}_{j\tilde{\varepsilon}}\mu_j > \frac{t^{1/2}n^{1/2}}{(2\varepsilon'\mathcal{M}_{s_0}\varepsilon)^{1/2}}$, summing over $j \in A_2$ gives

$$\sum_{j \in A_2} \varepsilon' H (H' H)^{-1} H' \frac{1}{(\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}} \mathcal{M}_{s_0} \tilde{w}_j \mu_j > m_2 \frac{t^{1/2} n^{1/2}}{(2\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}}$$

This implies that

$$\left\| (H' H)^{-1} H' \frac{1}{(\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}} \sum_{j \in A_2} \mathcal{M}_{s_0} \tilde{w}_j \mu_j \right\|_1 \|\varepsilon' H\|_\infty > m_2 \frac{t^{1/2} n^{1/2}}{(2\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}}$$

Which further implies that

$$\sqrt{m + s_0} \left\| (H' H)^{-1} H' \frac{1}{(\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}} \sum_{j \in A_2} \mathcal{M}_{s_0} \tilde{w}_j \mu_j \right\|_2 \|\varepsilon' H\|_\infty > m_2 \frac{t^{1/2} n^{1/2}}{(2\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}}$$

Next, further upper bound the $\|\cdot\|_2$ term on the left side above by

$$\begin{aligned} & \left\| (H' H)^{-1} H' \frac{1}{(\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}} \sum_{j \in A_2} \mathcal{M}_{s_0} \tilde{w}_j \mu_j \right\|_2 \\ & \leq \frac{n^{-1/2}}{(\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}} \varphi_{\min}(s_0 + m)(G)^{-1/2} \|\mathcal{M}_{s_0} \sum_{j \in A_2} \tilde{w}_j \mu_j\|_2 \end{aligned}$$

next, by the fact that \mathcal{M}_{s_0} is a projection (hence non-expansive) and \tilde{w}_j are mutually orthogonal,

$$\leq \frac{n^{-1/2}}{(\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}} \varphi_{\min}(s_0 + m)(G)^{-1/2} \sqrt{\sum_{j \in A_2} \|\tilde{w}_j \mu_j\|_2^2}.$$

In Section 3, it was shown that $\max_j \|\tilde{w}_j\|_2^2 \leq \varphi_{\min}(s_0 + m)(G)^{-1}$. Therefore, putting the above inequalities together,

$$\frac{n^{-1/2}}{(\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}} \sqrt{m + s_0} \varphi_{\min}(m + s_0)(G)^{-1} \sqrt{m_2} \|\varepsilon' H\|_\infty > m_2 \frac{t^{1/2} n^{1/2}}{(2\varepsilon' \mathcal{M}_{s_0} \varepsilon)^{1/2}}.$$

This implies that

$$\begin{aligned} m_2 & < \frac{1}{n} \frac{2}{t} (\varepsilon' \mathcal{M}_{s_0} \varepsilon)(m + s_0) \frac{\|\varepsilon' H\|_\infty^2}{\varepsilon' \mathcal{M}_{s_0} \varepsilon} \varphi_{\min}(m + s_0)(G)^{-2} \\ & \leq 2(m + s_0) \frac{\|\mathbb{E}_n[x_i \varepsilon_i]\|_\infty^2}{t} \varphi_{\min}(m + s_0)(G)^{-2}. \end{aligned}$$

Under the assumed condition that $t^{1/2} \geq 2\|\mathbb{E}_n[x_i \varepsilon_i]\|_\infty \varphi_{\min}(m + s_0)(G)^{-1}$, it follows that

$$m_2 \leq \frac{1}{2}(m + s_0).$$

By substituting $m = m_1 + m_2$ gives $m_2 \leq m_1 + s_0$.

Step 7

This section concludes the proof of the second statement of the theorem by bringing together all of the facts proven in Steps 3-6. Combining $m_1 \leq \varphi_{\min}(m + s_0)(G)^{-1}C_1^{-2}C_3^2 (K_G^{\mathbb{R}})^2 s_0$ and $m_2 \leq m_1 + s_0$ gives

$$m \leq \left[2\varphi_{\min}(m + s_0)(G)^{-1}C_1^{-2}C_3^2 K_G^{\mathbb{R}^2} + 1 \right] s_0.$$

In addition,

$$C_1 = \frac{1}{2}\varphi_{\min}(m + s_0)(G),$$

$$C_2 = \varphi_{\min}(m + s_0)(G),$$

$$C_3 = (1 + \varphi_{\min}(m + s_0)(G)^{-1/2} + \varphi_{\min}(m + s_0)(G)^{-1}),$$

and $K_G^{\mathbb{R}} < 1.783$. Therefore,

$$m \leq \left[1 + 8 \times 1.783^2 \times \varphi_{\min}(m + s_0)(G)^{-3} \right. \\ \left. \times (1 + \varphi_{\min}(m + s_0)(G)^{-1/2} + \varphi_{\min}(m + s_0)(G)^{-1})^2 \right] s_0.$$

Since $C_3^2 \leq 9\varphi_{\min}(s_0 + m)(G)^{-2}$, the expression above can be simplified at the expense of a slightly less tight constant, so that

$$m \leq [1 + 72 \times 1.783^2 \times \varphi_{\min}(m + s_0)(G)^{-5}]s_0.$$

Since this bound holds for each positive integer m of wrong selections, this concludes the proof of Theorem 1. □

Appendix B: Proof of Theorems 2 and 3

Theorem 2 follows by applying Theorem 1 in the following way. If \hat{s} grows faster than s_0 , then there is $m < \hat{s}$ such that $s_0 < m < K_n$ and m/s_0 exceeds $c'_F(K_n) = O(1)$, giving a contradiction. The first statement of the theorem follows from applying the bound on \hat{s} . Theorem 3 follows by $\|\theta_0 - \hat{\theta}\|_1 \leq \sqrt{\hat{s} + s_0}\|\theta_0 - \hat{\theta}\|_2 \leq \sqrt{\hat{s} + s_0}\varphi_{\min}(\hat{s} + s_0)(G)^{-1}\mathbb{E}_n[(x'_i\theta_0 - x'_i\hat{\theta})^2]^{1/2}$.

Appendix C: Proof of Theorem 4

Proof. The proof of Theorem 4 is similar to the proof of Theorem 1. The quantities $\Delta_j\mathcal{E}(S)$ replace $\Delta_j\ell(S)$. The constants C_1 and C_2 depend on $c_{\text{test}}, c'_{\text{test}}, c''_{\text{test}}$ instead of t . The steps that follow mirror those taken in the proof of Theorem 1.

Step 1.

Let \mathcal{T} be the event described by Condition 2. Then by assumption, $P(\mathcal{T}) \geq 1 - \alpha - 3\delta_{\text{test}}/3 = 1 - \alpha - \delta_{\text{test}}$. The rest of the proof works on the event \mathcal{T} .

Suppose that Algorithm 1 terminates before $K_{\text{test}} - s_0$ steps. Note that on \mathcal{T} , Algorithm 1 terminates at a step with $-\Delta_j \mathcal{E}(\widehat{S}) \leq c_{\text{test}}$ for every $j \notin \widehat{S}$. By application of the results of [27], 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 $|S_0 \setminus \widehat{S}| \leq s_0$ yields

$$|\mathcal{E}(S_0) - \mathcal{E}(\widehat{S})| \leq \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1} \sum_{j \in S_0 \setminus \widehat{S}} -\Delta_j \mathcal{E}(S) \leq s_0 c_{\text{test}} \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1}.$$

Step 2.

Repeat the construction in Step 2 of the proof of Theorem 1. The inner product on \mathbb{R}^n is replaced by the standard inner product on $L^2(\Omega; \mathbb{R}^n)$ where Ω is an underlying probability space. For example, $\langle x_j, x_k \rangle_{L^2(\Omega; \mathbb{R}^n)} = \mathbb{E}[\mathbb{E}_n[x_{ij}x_{ik}]]$. This yields parameters $\tilde{\theta}, \tilde{\gamma}_j \in \mathbb{R}^{s_0}$ and $\tilde{\theta}_{\tilde{\varepsilon}}, \tilde{\gamma}_{j\tilde{\varepsilon}} \in \mathbb{R}$ as well as an implied profile of operators $\{\mathcal{M}_k^{(L^2(\Omega; \mathbb{R}^n))}\}_{k=1}^{s_0}$ and $\{\mathcal{Q}_S^{(L^2(\Omega; \mathbb{R}^n))}\}_{S \subseteq \{1, \dots, p\}}$ on $L^2(\Omega; \mathbb{R}^n)$.

Step 3.

Divide the set of false covariates into three sets A_1 , A_2 , and A_3 where

$$A_1 = \left\{ j < K_{\text{test}} - s_0 : |\tilde{\gamma}_{j\tilde{\varepsilon}}| \leq \frac{c'_{\text{test}}{}^{1/2}}{(2\langle \varepsilon, \mathcal{M}_{s_0} \varepsilon \rangle_{L^2(\Omega; \mathbb{R}^n)})^{1/2}} \right\}$$

$$A_2 = \left\{ j < K_{\text{test}} - s_0 : |\tilde{\gamma}_{j\tilde{\varepsilon}}| > \frac{c'_{\text{test}}{}^{1/2}}{(2\langle \varepsilon, \mathcal{M}_{s_0} \varepsilon \rangle_{L^2(\Omega; \mathbb{R}^n)})^{1/2}} \right\}$$

$$A_3 = \{j : j \geq K_{\text{test}} - s_0\}.$$

By arguments identical to Step 3 in the proof of Theorem 1, for any $e_j \in \{-1, +1\}$,

$$\left\| \sum_{j \in A_1} e_j \tilde{\gamma}_j \right\|_2^2 \leq |A_1| \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1}.$$

Step 4.

Let C_1 be such that $\tilde{\theta}'\tilde{\gamma}_j/\tilde{\theta}_k > C_1$ for $j \in A_{1k}$. Let C_2 be such that $\tilde{\theta}_k/\tilde{\theta}_l > C_2$ for $l > k$. Let $C_3 = 1 + C_2^{-1} + C_2^{-1/2}$. By arguments identical to Step 4 in the proof of Theorem 1,

$$|A_1| \leq \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1}C_1^{-2}C_3^2(K_G^{\mathbb{R}})^2s_0.$$

Step 5.

This section calculates bounds C_1 and C_2 needed in Step 4 immediately above. Proceeding as in Step 5 of the proof of Theorem 1, note that

$$-\Delta_j \mathcal{E}(S_{\text{pre-}w_j}) = \langle y, \tilde{w}_j \rangle (\langle \tilde{w}_j, \tilde{w}_j \rangle)^{-1} \langle \tilde{w}_j, y \rangle.$$

This allows calculation of C_1 and C_2 in Theorem 3 to proceed in a similar manner as in Theorem 1. The statement $\Delta_k \ell(S_{\text{pre-}w_j}) \geq t$ is replaced by

$$\Delta_k \mathcal{E}(S_{\text{pre-}w_j}) \geq c'_{\text{test}}$$

for j selected. Similarly, the statement $\Delta_j \ell(S_{\text{pre-}w_j}) \geq \Delta_k \ell(S_{\text{pre-}w_j})$ is replaced by

$$\Delta_j \mathcal{E}(S_{\text{pre-}w_j}) \geq c''_{\text{test}} \Delta_k \mathcal{E}(S_{\text{pre-}w_j}).$$

Using similar reasoning

$$\tilde{\theta}'\tilde{\gamma}_j/\tilde{\theta}_k \geq \frac{1}{2}\varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])c''_{\text{test}} = C_1$$

and

$$\tilde{\theta}_k/\tilde{\theta}_l \geq \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])c''_{\text{test}} = C_2.$$

Step 6.

By similar arguments as in Step 6 in the proof of Theorem 1, $|A_1 \cup A_2| \leq 2|A_1| + s_0$. Note that relative to Theorem 1, some care must be taken to define what is meant by H . H is a vector of elements of $L_2(\Omega; \mathbb{R}^n)$ and H' is a vector of elements of the dual space $L_2(\Omega; \mathbb{R}^n)^*$.

Step 7.

By identical arguments as in Step 7 in the proof of Theorem 1,

$$|A_1 \cup A_2| \leq \left[2\varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1}C_1^{-2}C_3^2K_G^{\mathbb{R}^2} + 1 \right] s_0.$$

Using the values for C_1 and C_2 in Step 5,

$$|A_1 \cup A_2| \leq [1 + 72 \times 1.783^2 \times \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-5}c_{\text{test}}''^{-4}]s_0.$$

Finally, if $K_{\text{test}} - s_0$ is larger than $[1 + 72 \times 1.783^2 \times \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-5}c_{\text{test}}''^{-4}]s_0$, then it must follow that $|A_3| = 0$. Then $\hat{s} \leq |A_1 \cup A_2| + s_0$, and from this the result follows.

Step 8.

This step proves the third statement of Theorem 4. For any S define θ_S^* to be the minimizer of $\mathcal{E}(S)$. For any S define also $d_S = \theta_S^* - \theta_{S_0}^*$. Finally, let $\delta_0 = \theta_0 - \theta_{S_0}^*$. Note that $\mathcal{E}(S) - \mathcal{E}(S_0) = d_S' \mathbb{E}[G] d_S$. By the earlier steps, $d_S' \mathbb{E}[G] d_S \leq s_0 c_{\text{test}} \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1}$. But $d_S' \mathbb{E}[G] d_S \geq \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G]) \|d_S\|_2^2$. So $\|d_S\|_2 \leq \sqrt{s_0 c_{\text{test}} \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1}}$. In addition, δ_0 is bounded by

$$\begin{aligned} \|\delta_0\|_2 &= \|\mathbb{E}[G]^{-1} \mathbb{E}[\mathbb{E}_n[x_i x_i' \theta_0 + \varepsilon_i]] - \theta_0\|_2 = \|\mathbb{E}[\mathbb{E}_n[x_{iS_0}' \varepsilon_i]]\|_2 \\ &\leq s_0^{1/2} \max_j |\mathbb{E}[\mathbb{E}_n[x_{ij} \varepsilon_i^a]]| \leq \frac{1}{2} \sqrt{s_0 c_{\text{test}}} \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1} \end{aligned}$$

where the last bound comes from Cauchy-Schwarz (passing to $\mathbb{E}[\mathbb{E}_n[x_{ij}^2]]^{1/2} \mathbb{E}[\mathbb{E}_n[\varepsilon_i^{a2}]]^{1/2}$) along with the assumed condition on ε_i^a and the fact that $c_{\text{test}}' \leq c_{\text{test}}$. Next,

$$\begin{aligned} \hat{\theta} &= G_{\hat{S}}^{-1} \mathbb{E}_n[x_{i\hat{S}}(x_{i\hat{S}}' \theta_{\hat{S}}^* + \varepsilon_i - x_{i\hat{S} \cup S_0}' d_{\hat{S}} + x_{iS_0}' \delta_0)] \\ &= \theta_{\hat{S}}^* + G_{\hat{S}}^{-1} \mathbb{E}_n[x_{i\hat{S}} \varepsilon_i] + G_{\hat{S}}^{-1} \mathbb{E}_n[x_{i\hat{S}} x_{i\hat{S} \cup S_0}' (-d_{\hat{S}} + \delta_0)] \\ \Rightarrow \|\hat{\theta} - \theta_{\hat{S}}^*\|_2 &\leq \varphi_{\min}(\hat{s})(G)^{-1/2} \|\mathbb{E}_n[x_{i\hat{S}} \varepsilon_i]\|_2 + \left\| G_{\hat{S}}^{-1} \mathbb{E}_n[x_{i\hat{S}} x_{i\hat{S} \cup S_0}' (-d_{\hat{S}} + \delta_0)] \right\|_2 \\ &\leq \varphi_{\min}(\hat{s})(G)^{-1/2} \hat{s}^{1/2} \|\mathbb{E}_n[x_i \varepsilon_i]\|_{\infty} \\ &\quad + \varphi_{\min}(\hat{s})(G)^{-1/2} \varphi_{\max}(\hat{s} + s_0)(G)^{1/2} (\|d_{\hat{S}}\|_2 + \|\delta_0\|_2). \end{aligned}$$

Finally,

$$\begin{aligned}
(\mathbb{E}_n[(x'_i \widehat{\theta} - x'_i \theta_0)^2])^{1/2} &\leq \varphi_{\max}(s_0 + \widehat{s})(G)^{1/2} \|\widehat{\theta} - \theta_0\|_2 \\
&\leq \varphi_{\max}(s_0 + \widehat{s})(G)^{1/2} (\|\widehat{\theta} - \theta_{\widehat{s}}^*\|_2 + \|\delta_0\|_2 + \|d_{\widehat{s}}\|_2) \\
&\leq \varphi_{\max}(s_0 + \widehat{s})(G)^{1/2} \varphi_{\min}(s_0 + \widehat{s})(G)^{-1/2} \widehat{s}^{1/2} \|\mathbb{E}_n[x_i \varepsilon_i]\|_{\infty} \\
&\quad + \varphi_{\max}(s_0 + \widehat{s})(G)^{1/2} \left(\frac{3}{2} + \frac{3}{2}\right) \varphi_{\max}(s_0 + \widehat{s})(G)^{1/2} \varphi_{\min}(\widehat{s} + s_0)(G)^{-1/2} \\
&\quad \times \sqrt{s_0 c_{\text{test}}} \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1} \\
&\leq \varphi_{\max}(s_0 + \widehat{s})(G)^{1/2} \varphi_{\min}(s_0 + \widehat{s})(G)^{-1/2} \widehat{s}^{1/2} \|\mathbb{E}_n[x_i \varepsilon_i]\|_{\infty} \\
&\quad + 3\varphi_{\max}(s_0 + \widehat{s})(G) \varphi_{\min}(\widehat{s} + s_0)(G)^{-1/2} \sqrt{s_0 c_{\text{test}}} \varphi_{\min}(K_{\text{test}})(\mathbb{E}[G])^{-1}.
\end{aligned}$$

□

Appendix D: Proof of Theorem 5

Step 1.

This section sets up basic notation and outlines the proof. The strategy is to apply Theorem 4 using the conditional distribution P_x on for \mathcal{D}_n , conditional on x . The unconditional result is then shown to follow. Let $\mathcal{E}_x(S) = \mathbb{E}[\ell(S)|x]$. In addition, let $\theta_{jS}^{*|x} = (x_j \mathcal{Q}_S x_j)^{-1} \mathbb{E}[x_j \mathcal{Q}_S (x \theta_0 + \varepsilon^a)|x]$.

Next let

$$\widehat{Z}_{jS} = \widehat{V}_{jS}^{-1/2} (\widehat{\theta}_{jS} - [\theta_{jS}^{*|x}]_j).$$

Let $t_{\alpha} = \Phi^{-1}(1 - \alpha/p)$. Let \mathcal{A} be the event given by

$$\mathcal{A} = \left\{ |\widehat{Z}_{jS}| \leq \left(\frac{1 + c_{\tau}}{2}\right) \widehat{\tau}_{jS} t_{\alpha} \text{ for all } j, |S| < K_n \right\}.$$

Note that

$$\Delta_j \mathcal{E}_x(S) = [\theta_{jS}^{*|x}]_j^2 A_{jS}$$

for A_{jS} defined by $A_{jS} = [G_{jS}^{-1}]_{jj}$.

The next three steps calculate the extent to which the event \mathcal{A} implies that the tests have power, control size, and exhibit continuity. Step 2 performs a power calculation. Step 3 performs a size calculation, and Step 4 performs a continuity calculation for the test statistics.

Step 2.

This section calculates power properties. Suppose that

$$-\Delta_j \mathcal{E}_x(S) \geq A_{jS} \widehat{V}_{jS} (c_\tau + 1)^2 \widehat{\tau}_{jS}^2 t_\alpha^2.$$

Then on \mathcal{A} , and for $|S| < K_n$,

$$\begin{aligned} [\theta_{jS}^{*|x}]_j^2 A_{jS} &\geq A_{jS} \widehat{V}_{jS} (c_\tau + 1)^2 \widehat{\tau}_{jS}^2 t_\alpha^2 \\ |[\theta_{jS}^{*|x}]_j| &\geq \widehat{V}_{jS}^{1/2} (c_\tau + 1) \widehat{\tau}_{jS} t_\alpha \\ |[\widehat{\theta}_{jS}]_j| &\geq \widehat{V}_{jS}^{1/2} (c_\tau + 1) \widehat{\tau}_{jS} t_\alpha - |[\theta_{jS}^{*|x}]_j - [\widehat{\theta}_{jS}]_j| \\ |[\widehat{\theta}_{jS}]_j| &\geq \widehat{V}_{jS}^{1/2} (c_\tau + 1) \widehat{\tau}_{jS} t_\alpha - \widehat{V}_{jS}^{1/2} \left(\frac{1 + c_\tau}{2} \right) \widehat{\tau}_{jS} t_\alpha \\ |[\widehat{\theta}_{jS}]_j| &\geq \widehat{V}_{jS}^{1/2} c_\tau \widehat{\tau}_{jS} t_\alpha \end{aligned}$$

which implies $T_{jS\alpha} = 1$.

Step 3.

This step performs a size calculation. By construction, if $T_{jS\alpha} = 1$ then $|\widehat{V}_{jS}^{-1/2} [\widehat{\theta}_{jS}]_j| \geq c_\tau \widehat{\tau}_{jS} t_\alpha$, which is equivalent to

$$|[\widehat{\theta}_{jS}]_j| \geq c_\tau \widehat{\tau}_{jS} t_\alpha \widehat{V}_{jS}^{1/2}.$$

On \mathcal{A} , for $|S| < K_n$, note that

$$|[\widehat{\theta}_{jS}]_j - [\theta_{jS}^{*|x}]_j| \leq \widehat{V}_{jS}^{1/2} \left(\frac{1 + c_\tau}{2} \right) \widehat{\tau}_{jS} t_\alpha.$$

Then $T_{jS\alpha} = 1 \Rightarrow$

$$\begin{aligned} |[\theta_{jS}^{*|x}]_j| &\geq c_\tau \widehat{\tau}_{jS} t_\alpha \widehat{V}_{jS}^{1/2} - \widehat{V}_{jS}^{1/2} \left(\frac{1 + c_\tau}{2} \right) \widehat{\tau}_{jS} t_\alpha \\ &= \widehat{V}_{jS}^{1/2} \widehat{\tau}_{jS} t_\alpha \left(\frac{c_\tau - 1}{2} \right) \end{aligned}$$

Then

$$-\Delta_j \mathcal{E}_x(S) \geq A_{jS} \widehat{V}_{jS} \widehat{\tau}_{jS}^2 t_\alpha^2 \left(\frac{c_\tau - 1}{2} \right)^2.$$

Step 4.

This section calculates continuity properties of the tests. Suppose that $T_{jS\alpha} = T_{kS\alpha} = 1$, and that $W_{jS} \leq W_{kS}$. Note that $W_{jS} \leq W_{kS}$ implies $\widehat{V}_{jS}^{-1/2} |[\widehat{\theta}_{jS}]_j| \leq \widehat{V}_{kS}^{-1/2} |[\widehat{\theta}_{kS}]_k|$. Then on \mathcal{A} , and for $|S| < K_n$,

$$\widehat{V}_{jS}^{-1/2} |[\theta_{jS}^{*|x}]_j| - \left(\frac{1+c_\tau}{2} \right) \widehat{\tau}_{jS} t_\alpha \leq \widehat{V}_{kS}^{-1/2} |[\theta_{kS}^{*|x}]_k| + \left(\frac{1+c_\tau}{2} \right) \widehat{\tau}_{kS} t_\alpha$$

$$\widehat{V}_{jS}^{-1/2} |[\theta_{jS}^{*|x}]_j| \leq \widehat{V}_{kS}^{-1/2} |[\theta_{kS}^{*|x}]_k| + 2 \left(\frac{1+c_\tau}{2} \right) \widehat{\tau}_{kS} t_\alpha$$

$$\widehat{V}_{jS}^{-1/2} A_{jS}^{-1/2} (-\Delta_j \mathcal{E}_x(S))^{1/2} \leq \widehat{V}_{kS}^{-1/2} A_{kS}^{-1/2} (-\Delta_k \mathcal{E}_x(S))^{1/2} + 2 \left(\frac{1+c_\tau}{2} \right) \widehat{\tau}_{kS} t_\alpha$$

Using the fact that $-\Delta_k \mathcal{E}_x(S) \geq A_{kS} \widehat{V}_{kS} \left(\frac{1-c_\tau}{2} \right)^2 \widehat{\tau}_{kS}^2 t_\alpha^2$ derived in Step 2 and the fact that $T_{kS\alpha} = 1$, gives that $\left(\frac{1+c_\tau}{2} \right) \widehat{\tau}_{kS} t_\alpha \leq \widehat{V}_{kS}^{-1/2} A_{kS}^{-1/2} (-\Delta_k \mathcal{E}_x(S))^{1/2} \left(\frac{1+c_\tau}{c_\tau-1} \right)$.

Then

$$\widehat{V}_{jS}^{-1/2} A_{jS}^{-1/2} (-\Delta_j \mathcal{E}_x(S))^{1/2} \leq \widehat{V}_{kS}^{-1/2} A_{kS}^{-1/2} (-\Delta_k \mathcal{E}_x(S))^{1/2} + 2 \widehat{V}_{kS}^{-1/2} A_{kS}^{-1/2} (-\Delta_k \mathcal{E}_x(S))^{1/2} \left(\frac{1+c_\tau}{c_\tau-1} \right).$$

This gives

$$-\Delta_j \mathcal{E}_x(S) \leq \frac{\widehat{V}_{jS} A_{jS}}{\widehat{V}_{kS} A_{kS}} \left(1 + 2 \frac{1+c_\tau}{c_\tau-1} \right)^2 (-\Delta_k \mathcal{E}_x(S))$$

Step 5

This step summarizes Steps 2-4 and outlines the next steps. By Steps 2-4, the following implications are valid on \mathcal{A} for all $j, |S| < K_n$:

1. $T_{jS\alpha} = 1$ if $-\Delta_j \mathcal{E}_x(S) \geq A_{jS} \widehat{V}_{jS} (c_\tau + 1)^2 \widehat{\tau}_{jS}^2 t_\alpha^2$.
2. $-\Delta_j \mathcal{E}_x(S) \geq A_{jS} \widehat{V}_{jS} \left(\frac{1-c_\tau}{2} \right)^2 \widehat{\tau}_{jS}^2 t_\alpha^2$ if $T_{jS\alpha} = 1$.
3. $-\Delta_j \mathcal{E}_x(S) \leq \frac{\widehat{V}_{jS} A_{jS}}{\widehat{V}_{kS} A_{kS}} \left(1 + 2 \frac{1+c_\tau}{c_\tau-1} \right)^2 (-\Delta_k \mathcal{E}_x(S))$ if $T_{jS\alpha} = T_{kS\alpha} = 1, W_{jS} \leq W_{kS}$.

Next define a sequence of sets $\mathcal{X} = \mathcal{X}_n$ which will be shown to have the property that both $P(x \in \mathcal{X}) \rightarrow 1$ and

$$P^{\mathcal{X}}(\mathcal{A}) = \inf_{x \in \mathcal{X}} P(\mathcal{A}|x) \rightarrow 1.$$

In addition, there will be constants $\tilde{c}_{\text{test}}, \tilde{c}'_{\text{test}}, c''_{\text{test}} > 0$ which are independent of n and the realization of x , such that for $c_{\text{test}} = \frac{1}{n}\tilde{c}_{\text{test}}, c'_{\text{test}} = \frac{1}{n}\tilde{c}'_{\text{test}}$ and for the set \mathcal{B} defined by

$$\mathcal{B} = \begin{cases} 1. & A_{jS}\widehat{V}_{jS}(c_{\tau} + 1)^2\widehat{\tau}_{jS}^2 t_{\alpha}^2 \leq c_{\text{test}} \\ 2. & A_{jS}\widehat{V}_{jS}\left(\frac{1-c_{\tau}}{2}\right)^2\widehat{\tau}_{jS}^2 t_{\alpha}^2 \geq c'_{\text{test}} & |S| < K_n \\ 3. & \frac{A_{jS}\widehat{V}_{jS}}{A_{kS}\widehat{V}_{kS}}\left(1 + 2\frac{c_{\tau}-1}{c_{\tau}+1}\right)^2 \geq c''_{\text{test}} \end{cases}$$

it holds that $P^{\mathcal{X}}(\mathcal{B}) \rightarrow 1$.

Define sets $\mathcal{X} = \mathcal{X}_n$ as follows.

$$\mathcal{X} = \mathcal{X}_1 \cap \mathcal{X}_2 \cap \mathcal{X}_3 \cap \mathcal{X}_4$$

$$\mathcal{X}_1 = \{x : \max_{j \leq p} \mathbb{E}_n[x_{ij}^2] = O(1)\}$$

$$\mathcal{X}_2 = \{x : \varphi_{\min}(K_n)(G)^{-1} = O(1)\}$$

$$\mathcal{X}_3 = \{x : \|\eta_{jS}\|_1 = O(1)\}$$

$$\mathcal{X}_4 = \{x : P(\varphi_{\min}(K_n)(\mathbb{E}_n[\varepsilon_i^2 x_i x'_i])^{-1} = O(1)|x) = 1 - o(1)\}$$

Note that $P(\mathcal{X}_1), P(\mathcal{X}_2), P(\mathcal{X}_3) \rightarrow 1$ by assumption. In addition, failure of $P(\mathcal{X}_4) \rightarrow 1$ would contradict the unconditional statement in Condition 5 that $P(\varphi_{\min}(K_n)(\mathbb{E}_n[\varepsilon_i^2 x_i x'_i])^{-1} = O(1)) = 1 - o(1)$. Therefore, $P(\mathcal{X}) \rightarrow 1$.

Next, Step 6 proves that $P^{\mathcal{X}}(\mathcal{A}) \rightarrow 1$ and Step 7 proves that $P^{\mathcal{X}}(\mathcal{B}) \rightarrow 1$. Note that these steps allow the application of Theorem 4 conditionally on x . Step 8 uses this fact, and then concludes the proof by showing that θ_0 is bounded to $\theta_{S_0}^{*|x}$.

Step 6.

Note that

$$\begin{aligned}
\widehat{Z}_{jS} &= \widehat{V}_{jS}^{-1/2}([\widehat{\theta}_{jS}]_j - [\theta_{jS}^*]_j) \\
&= \widehat{V}_{jS}^{-1/2}(x_j' Q_S x_j)^{-1} x_j' Q_S (\varepsilon - \mathbf{E}[\varepsilon|x]) \\
&= ((x_j' Q_S x_j)^{-1} \mathbb{E}_n[\widehat{\varepsilon}_{ijS}^2 [Q_S x_{jS}]_i^2] (x_j' Q_S x_j)^{-1})^{-1} (x_j' Q_S x_j)^{-1} x_j' Q_S (\varepsilon - \mathbf{E}[\varepsilon|x]) \\
&= \mathbb{E}_n[\widehat{\varepsilon}_{ijS}^2 [Q_S x_{jS}]_i^2]^{-1} x_j' Q_S (\varepsilon - \mathbf{E}[\varepsilon|x]) \\
&= \mathbb{E}_n[\widehat{\varepsilon}_{ijS}^2 (\eta_{jS}' x_{ijS})^2]^{-1} \eta_{jS}' x_{jS} (\varepsilon - \mathbf{E}[\varepsilon|x]). \\
&= \mathbb{E}_n[\widehat{\varepsilon}_{ijS}^2 (\eta_{jS}' x_{ijS})^2]^{-1} \eta_{jS}' x_{jS} (\varepsilon^o + \varepsilon^a - \mathbf{E}[\varepsilon^a|x]).
\end{aligned}$$

Let $\ddot{\varepsilon} = \varepsilon^o + \varepsilon^a - \mathbf{E}[\varepsilon^a|x]$.

Define the *Regularization Event* by

$$\mathcal{R} = \left\{ \frac{|\sum_{i=1}^n x_{ik} \ddot{\varepsilon}_i|}{\sqrt{\sum_{i=1}^n x_{ik}^2 \ddot{\varepsilon}_i^2}} \leq t_\alpha \text{ for every } k \leq p \right\}$$

In addition, define the *Variability Domination Event* by

$$\mathcal{V} = \left\{ \sum_{i=1}^n x_{ik}^2 \ddot{\varepsilon}_i^2 \leq \left(\frac{1+c_\tau}{2}\right)^2 \sum_{i=1}^n x_{ik}^2 \widehat{\varepsilon}_{ijS}^2 \text{ for every } k \in jS, \text{ for every } |S| < K_n \right\}$$

The definition of the Regularization Event and the Variability Domination Event are useful since

$$\mathcal{R} \cap \mathcal{V} \Rightarrow \mathcal{A}.$$

To see this, note that on \mathcal{R} , the following inequality holds for any conformable vector ν :

$$\left(\sum_{i=1}^n \sum_{k \in jS} \nu_k x_{ik} \ddot{\varepsilon}_i \right)^2 \leq \left(t_\alpha \sum_{k \in jS} |\nu_k| \sqrt{\sum_{i=1}^n x_{ik}^2 \ddot{\varepsilon}_i^2} \right)^2$$

Furthermore, on \mathcal{V} , the previous expression can be further bounded by

$$\begin{aligned}
&\leq \left(\frac{1+c_\tau}{2}\right)^2 \left(t_\alpha \sum_{k \in jS} |\nu_k| \sqrt{\sum_{i=1}^n x_{ik}^2 \widehat{\varepsilon}_{ijS}^2}\right)^2 \\
&= \left(\frac{1+c_\tau}{2}\right)^2 \frac{\left(t_\alpha \sum_{k \in jS} |\nu_k| \sqrt{\sum_{i=1}^n x_{ik}^2 \widehat{\varepsilon}_{ijS}^2}\right)^2}{\sum_{i=1}^n \left(\sum_{k \in jS} \nu_k x_{ik}\right)^2 \widehat{\varepsilon}_{ijS}^2} \sum_{i=1}^n \left(\sum_{k \in jS} \nu_k x_{ik}\right)^2 \widehat{\varepsilon}_{ijS}^2 \\
&= \left(\frac{1+c_\tau}{2}\right)^2 t_\alpha^2 \frac{\|\nu' \text{Diag}(\Psi_{jS}^\widehat{\varepsilon})^{1/2}\|_1^2}{\nu' \Psi_{jS}^\widehat{\varepsilon} \nu} \sum_{i=1}^n \left(\sum_{k \in jS} \nu_k x_{ik}\right)^2 \widehat{\varepsilon}_{ijS}^2.
\end{aligned}$$

Specializing to the case that $\nu = \eta_{jS}$, and using $\widehat{\tau}_{jS} = \frac{\|\nu' \text{Diag}(\Psi_{jS}^\widehat{\varepsilon})^{1/2}\|_1}{\sqrt{\nu' \Psi_{jS}^\widehat{\varepsilon} \nu}}$ gives that

$$|\widehat{Z}_{jS}| \leq \frac{1+c_\tau}{2} \widehat{\tau}_{jS} t_\alpha \quad \text{on } \mathcal{R} \cap \mathcal{V}.$$

It is therefore sufficient to prove that \mathcal{R} and \mathcal{V} have probability $\rightarrow 1$ under P^X . $P^X(\mathcal{R}) \rightarrow 1$ follows immediately from the moderate deviation bounds for self-normalized sums given in [38]. For details on the application of this result, see [6].

Therefore, it is only left to show that $P^X(\mathcal{V}) \rightarrow 1$. Define $\varepsilon_{ijS} = y_i - x'_{ijS} \theta_{jS}^{*|x}$. Furthermore, define ξ_{ijS} through the decomposition $\varepsilon_{ijS} = \ddot{\varepsilon}_i + \xi_{ijS}$. Let ε_{jS} and ξ_{jS} be the respective stacked versions. Let $\tilde{c}_\tau = ((1+c_\tau)/2)^2$.

Then

$$\begin{aligned}
\tilde{c}_\tau \sum_{i=1}^n x_{ik}^2 \widehat{\varepsilon}_{ijS}^2 &= \tilde{c}_\tau \left[\sum_{i=1}^n x_{ik}^2 (\widehat{\varepsilon}_{ijS}^2 - \varepsilon_{ijS}^2) + \sum_{i=1}^n x_{ik}^2 \ddot{\varepsilon}_i^2 + 2 \sum_{i=1}^n x_{ik}^2 \ddot{\varepsilon}_i \xi_{ijS} + \sum_{i=1}^n x_{ik}^2 \xi_{ijS}^2 \right] \\
&\geq \tilde{c}_\tau \left[\sum_{i=1}^n x_{ik}^2 (\widehat{\varepsilon}_{ijS}^2 - \varepsilon_{ijS}^2) + \sum_{i=1}^n x_{ik}^2 \ddot{\varepsilon}_i^2 + 2 \sum_{i=1}^n x_{ik}^2 \ddot{\varepsilon}_i \xi_{ijS} \right] \\
&= \sum_{i=1}^n x_{ik}^2 \ddot{\varepsilon}_i^2 \\
&\quad + \tilde{c}_\tau \sum_{i=1}^n x_{ik}^2 (\widehat{\varepsilon}_{ijS}^2 - \varepsilon_{ijS}^2) + \frac{(\tilde{c}_\tau - 1)}{2} \sum_{i=1}^n x_{ik}^2 \ddot{\varepsilon}_i^2 \\
&\quad + 2\tilde{c}_\tau \sum_{i=1}^n x_{ik}^2 \ddot{\varepsilon}_i \xi_{ijS} + \frac{(\tilde{c}_\tau - 1)}{2} \sum_{i=1}^n x_{ik}^2 \ddot{\varepsilon}_i^2.
\end{aligned}$$

Define the two events

$$\mathcal{V}' = \left\{ \tilde{c}_\tau \mathbb{E}_n[x_{ik}^2(\hat{\varepsilon}_{ijS}^2 - \varepsilon_{ijS}^2)] + \frac{(\tilde{c}_\tau - 1)}{2} \mathbb{E}_n[x_{ik}^2 \hat{\varepsilon}_i^2] \geq 0 \text{ for all } j, k \leq p, |S| < K_n \right\}$$

$$\mathcal{V}'' = \left\{ 2\tilde{c}_\tau \mathbb{E}_n[x_{ik}^2 \hat{\varepsilon}_i \xi_{ijS}] + \frac{(\tilde{c}_\tau - 1)}{2} \mathbb{E}_n[x_{ik}^2 \hat{\varepsilon}_i^2] \geq 0 \text{ for all } j, k \leq p, |S| < K_n \right\}$$

Therefore $\mathcal{V}' \cap \mathcal{V}'' \Rightarrow \mathcal{V}$.

Note that $\mathbb{E}_n[x_{ik}^2 \hat{\varepsilon}_i^2] \geq \frac{1}{2} \mathbb{E}_n[x_{ik}^2 \varepsilon_i^2] - \mathbb{E}_n[x_{ik}^2 \mathbb{E}[\varepsilon_i^a | x]] \geq \frac{1}{2} \mathbb{E}_n[x_{ik}^2 \varepsilon_i^2] - \max_{i \leq n} \mathbb{E}[\varepsilon_i^{a2} | x]^{1/2} \mathbb{E}_n[x_{ik}^4]^{1/2}$. This is bounded below with $\mathbb{P}^x \rightarrow 1$ by a positive constant independent of n . Therefore, to show that $\mathbb{P}^x(\mathcal{V}') \rightarrow 1$, $\mathbb{P}^x(\mathcal{V}'') \rightarrow 1$, it suffices to show $\mathbb{E}_n[x_{ik}^2(\hat{\varepsilon}_{ijS}^2 - \varepsilon_{ijS}^2)]$ and $\mathbb{E}_n[x_{ik}^2 \hat{\varepsilon}_i \xi_{ijS}]$ respectively are suitably smaller order.

First consider $\mathbb{E}_n[x_{ik}^2(\hat{\varepsilon}_{ijS}^2 - \varepsilon_{ijS}^2)]$. It is convenient to bound the slightly more general sum $\mathbb{E}_n[x_{ik} x_{il}(\hat{\varepsilon}_{ijS}^2 - \varepsilon_{ijS}^2)]$, since this will show up again in Step 8.

$$\begin{aligned} \mathbb{E}_n[x_{ik} x_{il}(\hat{\varepsilon}_{ijS}^2 - \varepsilon_{ijS}^2)] &= 2\mathbb{E}_n \left[x_{ik} x_{il} \varepsilon_{ijS} x'_{ijS} (\theta_{jS}^{*|x} - \hat{\theta}_{jS}) \right] + \mathbb{E}_n \left[x_{ik} x_{il} (x'_{ijS} (\theta_{jS}^{*|x} - \hat{\theta}_{jS}))^2 \right] \\ &\leq 2 \|\mathbb{E}_n[x_{ik} x_{il} \varepsilon_{ijS} x'_{ijS}]\|_2 \|\theta_{jS}^{*|x} - \hat{\theta}_{jS}\|_2 + \lambda_{\max} \mathbb{E}_n[x_{ik} x_{il} x_{ijS} x'_{ijS}] \|\theta_{jS}^{*|x} - \hat{\theta}_{jS}\|_2^2 \end{aligned}$$

Standard reasoning gives that $\|\theta_{jS}^{*|x} - \hat{\theta}_{jS}\|_2 \leq \varphi_{\min}(K_n)(G)^{-1/2} \sqrt{K_n} \|\mathbb{E}_n x_{ijS} \varepsilon_{ijS}\|_\infty$. Therefore, the bound continues

$$\begin{aligned} &\leq 2 \|\mathbb{E}_n[x_{ik} x_{il} \varepsilon_{ijS} x'_{ijS}]\|_2 \varphi_{\min}(K_n)(G)^{-1/2} \sqrt{K_n} \|\mathbb{E}_n x_{ijS} \varepsilon_{ijS}\|_\infty \\ &\quad + \lambda_{\max} \mathbb{E}_n[x_{ik} x_{il} x_{ijS} x'_{ijS}] \varphi_{\min}(K_n)(G)^{-1} K_n \|\mathbb{E}_n x_{ijS} \varepsilon_{ijS}\|_\infty^2. \end{aligned}$$

Note that $\lambda_{\max} \mathbb{E}_n[x_{ik} x_{il} x_{ijS} x'_{ijS}] \leq K_n \max_{j \leq p} \mathbb{E}_n[x_{ij}^4]$.

$$\begin{aligned} &\leq 2 \|\mathbb{E}_n[x_{ik} x_{il} \varepsilon_{ijS} x'_{ijS}]\|_2 \varphi_{\min}(K_n)(G)^{-1/2} \sqrt{K_n} \|\mathbb{E}_n x_{ijS} \varepsilon_{ijS}\|_\infty \\ &\quad + K_n^2 \max_{j \leq p} \mathbb{E}_n[x_{ij}^4] \varphi_{\min}(K_n)(G)^{-1} \|\mathbb{E}_n x_{ijS} \varepsilon_{ijS}\|_\infty^2 \end{aligned}$$

An application of Cauchy-Schwarz to the top line gives

$$\begin{aligned} &\leq 2\sqrt{K_n} \max_j \mathbb{E}_n[x_{ik}^4]^{1/2} \max_{j,S} \mathbb{E}_n[\varepsilon_{ijS}^2 x_{ij}^2]^{1/2} \varphi_{\min}(K_n)(G)^{-1/2} \sqrt{K_n} \|\mathbb{E}_n x_{ijS} \varepsilon_{ijS}\|_\infty \\ &\quad + K_n^2 \max_{j \leq p} \mathbb{E}_n[x_{ij}^4] \varphi_{\min}(K_n)(G)^{-1} \|\mathbb{E}_n x_{ijS} \varepsilon_{ijS}\|_\infty^2 \end{aligned}$$

Next, $\|\mathbb{E}_n x_{ijS} \varepsilon_{ijS}\|_\infty$ and $\mathbb{E}_n[\varepsilon_{ijS}^2 x_{ij}^2]^{1/2}$ are bounded using $\varepsilon_{ijS} = \varepsilon_i - \mathbb{E}[\varepsilon_i | x] + \xi_{ijS}$. Note that by construction $\|\mathbb{E}_n[x_{ijS} \xi_{ijS}]\|_\infty = 0$. Then

$$\|\mathbb{E}_n[x_{ijS} \varepsilon_{ijS}]\|_\infty \leq \|\mathbb{E}_n[x_i \varepsilon_i]\|_\infty + \|\mathbb{E}_n[x_i \mathbb{E}[\varepsilon_i^a | x]]\|_\infty$$

$$\leq \|\mathbb{E}_n[x_i \varepsilon_i]\|_\infty + \max_{j \leq p} \mathbb{E}_n[x_{ij}^2]^{1/2} \mathbb{E}_n[\mathbb{E}[\varepsilon_i^a | x]^2]^{1/2} = O(\sqrt{\log p/n})$$

with $\mathbb{P}^X \rightarrow 1$. Next,

$$\begin{aligned} \mathbb{E}_n[\varepsilon_{ijS}^2 x_{ij}^2] &\leq 3\mathbb{E}_n[\varepsilon_i^2 x_{ij}^2] + 3\mathbb{E}_n[\mathbb{E}[\varepsilon_i^{a2} | x] x_{ij}^2] + 3\mathbb{E}_n[\xi_{ijS}^2 x_{ij}^2] \\ &\leq 3\mathbb{E}_n[\varepsilon_i^2 x_{ij}^2] + 3\mathbb{E}_n[x_{ij}^2] \max_{i \leq n} \mathbb{E}[\varepsilon_i^{a2} | x] + 3\mathbb{E}_n[\xi_{ijS}^4]^{1/2} \mathbb{E}_n[x_{ij}^4]^{1/2} \end{aligned}$$

Next, $(\mathbb{E}_n[\xi_{ijS}^4])^{1/2} \leq O(1)s_0^2$ on $\mathcal{X}_1 \cap \mathcal{X}_3$. To see this, note $\xi_{jS} = Q_{jS} x \theta_0 = \sum_{l=1}^{s_0} Q_{jS} x_l \theta_{0,l} = \sum_{l=1}^{s_0} \eta_{l,(jS)} x_{ljS} = \tilde{\eta}_{jS} x_{S_0 \cup jS}$ for some new linear combination $\tilde{\eta}_{jS}$. Note that $\|\tilde{\eta}_{jS}\|_1 \leq s_0 O(1)$. Then $(\mathbb{E}_n[\xi_{ijS}^4])^{1/4} \leq \|\tilde{\eta}_{jS}\|_1 \max_{k \leq p} \mathbb{E}_n[x_{ik}^4]^{1/4}$ from which the bound follows.

Next consider $\mathbb{E}_n[x_{ik}^2 \tilde{\varepsilon}_i \xi_{ijS}]$. Consider two cases. In Case 1,

$$\mathbb{E}_n[x_{ik}^4 \xi_{ijS}^2]^{1/2} \leq \frac{1}{2\tilde{c}_\tau} \frac{(\tilde{c}_\tau - 1)}{2} \frac{\mathbb{E}_n[x_{ik}^2 \tilde{\varepsilon}_i^2]}{\mathbb{E}_n[\tilde{\varepsilon}_i^2]^{1/2}}.$$

In this case, $2\tilde{c}_\tau \mathbb{E}_n[x_{ik}^2 \tilde{\varepsilon}_i \xi_{ijS}] \leq \mathbb{E}_n[x_{ik}^4 \xi_{ijS}^2]^{1/2} \mathbb{E}_n[\tilde{\varepsilon}_i^2]^{1/2} \leq \frac{\tilde{c}_\tau - 1}{2}$, and the requirement of \mathcal{V}'' for k, j, S holds.

For Case 2, suppose the alternative that $\mathbb{E}_n[x_{ik}^4 x_{ijS}^2] > \frac{1}{2\tilde{c}_\tau} \frac{(\tilde{c}_\tau - 1)}{2} \frac{\mathbb{E}_n[x_{ik}^2 \tilde{\varepsilon}_i^2]}{\mathbb{E}_n[\tilde{\varepsilon}_i^2]^{1/2}}$ holds. Then $\mathbb{E}[\mathbb{E}_n[x_{ik}^4 \xi_{ijS}^2 \tilde{\varepsilon}_i^2] | x]$ is bounded away from zero by conditions on $\mathbb{E}[\varepsilon_i^2 | x]$ and $\max_i |\varepsilon_i^a|$. In addition, $\mathbb{E}[\mathbb{E}_n[|x_{ik}|^6 |\xi_{ijS}|^3 |\tilde{\varepsilon}_i|^3] | x] \leq \max_i \mathbb{E}[\tilde{\varepsilon}_i^3 | x] \mathbb{E}_n[|x_{ik}|^6 |\xi_{ijS}|^3] \leq O(1) \mathbb{E}_n[|x_{ik}|^6 |\xi_{ijS}|^3]$. This term is further bounded by

$$O(1) \mathbb{E}_n[x_{ik}^{12}]^{1/2} \mathbb{E}_n[|\xi_{ijS}|^6]^{1/2}$$

Using the same reasoning as bounding $\mathbb{E}_n[\xi_{ijS}^4]$ earlier, it follows that $\mathbb{E}_n[|\xi_{ijS}|^6]^{1/2} = O(1)s_0^3$. In addition, $\mathbb{E}_n[x_{ik}^{12}] = O(1)$. As a result, for those k, j, S which fall in Case 2, the self-normalized sum

$$= \max_{j,k,S \in \text{Case 2}} \frac{\sqrt{n} |\mathbb{E}_n[x_{ik}^2 \xi_{ijS} \tilde{\varepsilon}_i]|}{\sqrt{\mathbb{E}_n[x_{ik}^4 \xi_{ijS}^2 \tilde{\varepsilon}_i^2]}}$$

is $O(\log(p^{K_n}))$ with probability $1 - o(1)$ provided $\sqrt{\log(p^{K_n})} = o(n^{1/6}/(s_0^3)^{1/3})$. This holds under the assumed rate conditions. Then $\max_{j,k,S} |\mathbb{E}_n[x_{ik}^2 \xi_{ijS} \tilde{\varepsilon}_i]|$ is bounded by $\frac{1}{\sqrt{n}} O(\log(p^{K_n}) \max_{j,k,S} \sqrt{\mathbb{E}_n[x_{ik}^4 \xi_{ijS}^2 \tilde{\varepsilon}_i^2]})$. Furthermore, $\mathbb{E}_n[x_{ik}^4 \xi_{ijS}^2 \tilde{\varepsilon}_i^2] \leq \mathbb{E}_n[x_{ik}^8 \xi_{ijS}^4]^{1/2} \mathbb{E}_n[\tilde{\varepsilon}_i^4]^{1/2} \leq (\mathbb{E}_n[x_{ik}^{12}]^{2/3} \mathbb{E}_n[\xi_{ijS}^{12}]^{1/3})^{1/2} \mathbb{E}_n[\tilde{\varepsilon}_i^4]^{1/2} \leq O(1)s_0^2 \mathbb{E}_n[\tilde{\varepsilon}_i^4]^{1/2}$. Note that $\mathbb{E}_n[\tilde{\varepsilon}_i^4]^{1/2} \leq O(1)$ with $\mathbb{P}^X \rightarrow 1$. Together, these give that $\max_{j,k,S} \mathbb{E}_n[x_{ik}^2 \tilde{\varepsilon}_i \xi_{ijS}] = o(1)$ with $\mathbb{P}^X \rightarrow 1$. Finally, $\mathbb{P}^X(\mathcal{V}) \rightarrow 1$.

Step 7.

This step shows that $P^{\mathcal{X}}(\mathcal{B}) \rightarrow 1$. First, A_{jS} depend only on x and are bounded above and below by constants which do not depend on n on \mathcal{X} from the assumption on the sparse eigenvalues of G . For bounding $\widehat{\tau}_{jS}$ above and away from zero, since $1 \leq \|\eta_{jS}\|_1, \|\eta_{jS}\|_2 \leq O(1)$ on \mathcal{X} , it is sufficient to show that the eigenvalues of $\Psi_{jS}^{\widehat{\varepsilon}} = \mathbb{E}_n[x_{ijS}x'_{ijS}\widehat{\varepsilon}_{ijS}^2]$ remain bounded above and away from zero and that the diagonal terms of $\Psi_{jS}^{\widehat{\varepsilon}}$ remain bounded above and away from zero. Note that by arguments in last step (Step 6), it was shown that $\mathbb{E}_n[x_{ik}x_{il}(\widehat{\varepsilon}_{ijS} - \varepsilon_{ijS})] = O(\sqrt{\log p/n})$ with $P^{\mathcal{X}} \rightarrow 1$. Therefore, $\|\mathbb{E}_n[x_{ijS}x'_{ijS}\widehat{\varepsilon}_{ijS}^2] - \mathbb{E}_n[x_{ijS}x'_{ijS}\varepsilon_{ijS}^2]\|_{\mathcal{F}} = O(K_n\sqrt{\log p/n})$ with $P^{\mathcal{X}} \rightarrow 1$. Here, \mathcal{F} is the Frobenius norm. By the assumed rate condition, the above quantity therefore vanishes with $P^{\mathcal{X}} \rightarrow 1$.

Next,

$$\begin{aligned} \mathbb{E}_n[x_{ijS}x'_{ijS}\varepsilon_{ijS}^2] &= \mathbb{E}_n[x_{ijS}x'_{ijS}\varepsilon_i^2] + 2\mathbb{E}_n[x_{ijS}x'_{ijS}\varepsilon_i(\xi_{ijS} + \mathbb{E}[\varepsilon_i^a|x])] \\ &\quad + \mathbb{E}_n[x_{ijS}x'_{ijS}(\xi_{ijS} + \mathbb{E}[\varepsilon_i^a|x])^2] \end{aligned}$$

The first term above, $\mathbb{E}_n[x_{ijS}x'_{ijS}\varepsilon_i^2]$, has eigenvalues bounded away from zero for all j, S with $P^{\mathcal{X}} \rightarrow 1$. The third term above, $\mathbb{E}_n[x_{ijS}x'_{ijS}(\xi_{ijS} + \mathbb{E}[\varepsilon_i^a|x])^2]$ is positive semidefinite by construction. The second term above has Frobenius norm tending to zero for all j, S with $P^{\mathcal{X}} \rightarrow 1$. This, in conjunction with the fact that the eigenvalues of $\mathbb{E}_n[x_{ijS}x'_{ijS}\widehat{\varepsilon}_{ijS}]$ are bounded above and away from zero with $P^{\mathcal{X}} \rightarrow 1$ shows that the eigenvalues of $\Psi_{jS}^{\widehat{\varepsilon}} = \mathbb{E}_n[x_{ijS}x'_{ijS}\widehat{\varepsilon}_{ijS}^2]$ are bounded above and away from zero with $P^{\mathcal{X}} \rightarrow 1$. Finally, for bounding \widehat{V}_{jS} , it is sufficient to show that $\max_{k \leq p} \mathbb{E}_n[\varepsilon_i^2(\eta'_{jS}x_{ijS})^2]$ be bounded above. This follows immediately from $\mathbb{E}[\varepsilon_i^4|x]$ being uniformly bounded and $\max_{j,S} \|\eta_{jS}\|_1 = O(1)$ and $\max_{k \leq p} \mathbb{E}_n[x_{ik}^4] = O(1)$. These imply that $P^{\mathcal{X}}(\mathcal{B}) \rightarrow 1$.

Step 8

The previous steps show that Theorem 4 can be applied conditionally on x . Note that renormalizing the covariates to satisfy $\mathbb{E}_n[x_{ij}^2] = 1$ does not affect $\mathcal{E}_x(S)$ and therefore does not affect the conclusion of Step 5. Moreover, on \mathcal{X} , renormalizing does not effect boundedness of sparse eigenvalues of G . Therefore, by application of Theorem 4,

$$P^{\mathcal{X}} \left(\mathbb{E}_n[(x'_i\theta_0^x - x_i\widehat{\theta})^2]^{1/2} \leq O(\sqrt{s_0 \log p/n}) \right) \rightarrow 1.$$

To show Theorem 5, it is left to show that

$$P^{\mathcal{X}} \left(\mathbb{E}_n[(x'_i\theta_0^x - x_i\theta_0)^2]^{1/2} \leq O(\sqrt{s_0 \log p/n}) \right) \rightarrow 1.$$

This follows from assumptions on ε^a . Note that

$$\theta_0^{|x} - \theta_0 = (x'_{S_0} x_{S_0})^{-1} x'_{S_0} \mathbb{E}[\varepsilon^a | x].$$

As a result,

$$\begin{aligned} \|\theta_0 - \theta_0^{|x}\|_2 &\leq \varphi_{\min}(s_0)(G)^{-1/2} \|\mathbb{E}_n[x_{i_{S_0}} \mathbb{E}[\varepsilon_i^a | x]]\|_2 \\ &\leq \varphi_{\min}(s_0)(G)^{-1/2} \sqrt{s_0} \|\mathbb{E}_n[x_{ij} \mathbb{E}[\varepsilon_i^a | x]]\|_\infty. \end{aligned}$$

By the assumed rate conditions and by $\max_i \mathbb{E}[\varepsilon_i^a] = O(n^{-1/2})$, the bound follows.

Appendix E: Proof of Theorems 7 and 8

Proof. Theorem 7 follows by applying Theorem 4 in [6]. Theorem 8 follows by applying Theorem 2 of [14]. \square

Appendix F: Additional computation details

Computation of the simulation studies and empirical example are performed using the software Matlab R2015a. The simulation study uses in addition the package `Glmnet` for Matlab ([55]). Replication files as well as detailed instructions are available from the author.