Analysis of 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 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, which depend on the quality of the tests, for prediction error and the number of selected covariates. As an example, the bounds are then specialized to a case with heteroskedastic data, with tests constructed with the help of Huber-Eicker-White standard errors. Under the assumed regularity conditions, these tests lead to estimation convergence rates matching other common high-dimensional estimators including Lasso.

MSC 2010 subject classifications: 62J05, 62J07, 62L12.
JEL subject classifications: C55.
Keywords and phrases: model selection, forward regression, sparsity, hypothesis testing.

* First version: November 1, 2015. This version March 26, 2020. I gratefully acknowledge helpful discussions with Christian Hansen, Tim Conley, Kelly Reeve, Dan Zou, Nicolai Meinshausen, Marloes Maathius, Rahul Mazumder, Ryan Tibshirani, Trevor Hastie, Martin Schonger, Pietro Biroli, Michael Wolf, Lorenzo Casaburi, Hannes Schwandt, Ralph Ossa, Rainer Winkelmann, attendants at the ETH Zürich Seminar für Statistik Research Seminar, attendants at the Center for Law and Economics Internal Seminar, attendants at the Toulouse School of Economics Econometrics Seminar, research assistants Vincent Lohmann and Alexandre Jenni for assisting in several stages of verification of formal arguments, Yuming Pan for assistance with computation, as well as financial support from UZH and from the ETH Fellowship program. The content of this paper draws from two separate working papers posted on ArXiv (https://arxiv.org/abs/1702.01000 (Kozbur (2017a)) and https://arxiv.org/abs/1512.02666 (the current paper)). The projects have been merged in preparation for the publication process. A mechanical statement of a special case of TBFMS appears in Kozbur (2017b), which neither claims nor derives any theoretical results. This paper previously had the title "Testing-Based Forward Model Selection," which is also the title of Kozbur (2017b).

1. Introduction

This paper analyzes a procedure called *Testing-Based Forward Model Selection* (TBFMS) for high-dimensional econometric problems, which are 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.

The primary settings of this paper are high-dimensional sparse linear regression models, in which the number of covariates is allowed to be commensurate with or 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. 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. Such methods in the statistics literature date back to at least Efroymson (1966).

In practice, deciding which covariate gives the best additional predictive power relative to a working model is complicated by the fact that outcomes are observed with noise or are partly idiosyncratic. For example, in linear regression, a variable associated with 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. 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

[1] High-dimensional data may arise in several ways—data may be intrinsically high-dimensional with many characteristics per observation, or alternatively, researchers may obtain a large final set of covariates through forming interactions and transformations of underlying covariates.

[2] Simple 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. "Model" is used in the name of TBFMS to avoid confusion with sample selection problems common in econometrics.

model is then based on the largest value of an associated test statistic. Note that in this context, the hypothesis tests are solely serving a role as assisting model selection, not ex post inference.

There are several earlier analyses of Simple Forward Selection.[3] Wang (2009) 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. Zhang (2009) and Tropp (2004) prove performance bounds for greedy algorithms under a strong irrepresentability condition, which restricts the empirical covariance matrix of the predictors. Das and Kempe (2011) prove bounds on the relative performance in population R-squared of 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 Zhang (2009)).

As a preliminary, this paper proves new bounds on the predictive performance and number of selected covariates for Simple Forward Selection. The conditions required here are weaker that those used in Zhang (2009) and Tropp (2004) and impose no β-min restrictions or irrepresentability. The convergence rates here are most similar to the analysis of a Forward-Backward Selection in Zhang (2011), but require markedly different analysis since there is no chance to correct "over-selection mistakes."

This paper then gives performance bounds for TBFMS which depend directly on the quality of the profile of tests considered, as measured by five 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.

There are many other sensible approaches to high-dimensional estimation. An important and common approach to generic high-dimensional estimation problems is the Lasso. The Lasso minimizes a least squares criterion augmented with a penalty proportional to the ℓ_1 norm of the coefficient vector. For theoretical and simulation results for Lasso, see Frank and Friedman (1993) Tibshirani (1996), Hastie et al. (2009) Candès and Tao (2007) Bai and Ng (2008), Bickel et al. (2009), Huang et al. (2010), Bühlmann and van de Geer (2011), among many more. Other related methods include boosting (see Freund and Schapire (1996), Bühlmann (2006), Luo and Spindler (2017)), Least Angle Regression (see Efron et al. (2004)), Post-Lasso (see Belloni et al. (2012)), and many others. A recent related paper, Chudik et al. (2018), considers a different iterative model selection procedure which also involves using hypothesis tests. In Chudik et al. (2018), in the first iteration, a marginal regression of the outcome on each potential covariate is run. Once all marginal regressions are run, all significant covariates are included into a working

[3] TBFMS using different tests than proposed here is natively programmed in some statistical software, including SPSS, but is not previously formally justified in high-dimensional settings.

model. Each subsequent iteration works similarly.

The asymptotic estimation rates calculated here for TBFMS, applied to a constructed profile of tests for heteroskedastic data, match those standard for Lasso and Post-Lasso. Relative to the analysis of asymptotic properties of Lasso and related high-dimensional estimation techniques, analysis of TBFMS is complicated by the fact that the procedure is not the optimizer of a simple objective function. As a result, the theory also departs from the literature on m-estimation in a fundamental way.

A recent paper, Hastie et al. (2017), performs a systematic simulation analysis of statistical and computational performance of simple forward selection as well as a few additional estimators, including Lasso and best subset selection; see Bertsimas et al. (2016). The paper reports that in regression models with higher signal-to-noise ratios, forward selection performs favorably relative to Lasso; a finding consistent across many simulation settings.

This paper complements recent literature on sequential testing (see G'Sell et al. (2016), Li and Barber (2017), Tibshirani et al. (2014), Fithian et al. (2015)). Sequential testing 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 over such sequences of hypothesis tests. While the current paper focuses on statistical properties of estimates after TBFMS given properties of the implemented tests, future work may potentially combine the two types of problems.

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 Belloni et al. (2012)). Another example is the selection of a conditioning set to properly control for omitted variables bias when there are many control variables (see Zhang and Zhang (2014), Belloni et al. (2014), van de Geer et al. (2014), and Javanmard and Montanari (2014)). Bounds about the quality of the selected model are used to derive results about the quality of post-model selection estimation and to guide subsequent inference. Such applications require a model selection procedure with hybrid objectives: (1) produce a good fit, and (2) return a sparse set of variables. This paper addresses both objectives by providing sparsity and fit bounds for TBFMS.

In terms of computing, one fast implementation of forward selection depends on what is sometimes referred to as a "guided QR decomposition." Formally, simple forward selection can be computed in $O(npk)$ flops, with n being sample size, p being number of covariates, and k being number of steps (see for example Hastie et al. (2017)), and requires the storage of the QR decomposition of at most k variables. The version of TBFMS presented in the paper for data with heteroskedastic disturbances can be computed with the same order of time and storage requirements.[4]

[4] A modification of the least angle regression algorithm can be made to implement a similarly efficient computation of Lasso (see Efron et al. (2004), though this may require potentially more iterations, as covariates can multiple times both enter and exit a suitably defined active set which terminates as the selected set).

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. 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.

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 consist of a set of covariates $x_i \in \mathbb{R}^p$, as well as outcome variables $y_i \in \mathbb{R}$ for each observation $i = 1, ..., n$. 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 $s_0 < n$. (Below, exact orthogonality between ε_i and x_i will not be required; rather a notion of approximate orthogonality will be used. As a result, the framework also handles notions of approximate sparsity).

Define an empirical 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 for subsets $S \subseteq \{1, ..., p\}$

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

The estimation strategy proceeds by first searching for a sparse subset $\widehat{S} \subseteq \{1, ..., p\}$, 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 $i = 1, ..., n$.

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

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

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

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

and $-\Delta_j \ell(S) \geqslant -\Delta_k \ell(S)$ 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 $\widehat{S} = \varnothing$
For $1 \leqslant k \leqslant p$
  If $-\Delta_j \ell(S) > t$ for some $j \in \{1, ..., p\} \setminus \widehat{S}$
    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 $\widehat{\theta} \in \arg\min_{\theta : \mathrm{supp}(\theta) \subset \widehat{S}} \ell(\theta)$.

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

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

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

$$c_{\mathrm{F}}(\widehat{s}) = (\widehat{s} + s_0)^{1/2} \varphi_{\min}(\widehat{s} + s_0)(G)^{-1/2} \left[2 \| \mathbb{E}_n[x_i \varepsilon_i] \|_\infty + t^{1/2} \right].$$

Finally, for each positive integer m, let

$$c'_{\mathrm{F}}(m) = 80 \times \varphi_{\min}(m + s_0)(G)^{-4}.$$

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 $j \leqslant p$. Then under Algorithm 1 with threshold t,*

$$\mathbb{E}_n[(x_i'\theta_0 - x_i'\widehat{\theta})^2]^{1/2} \leqslant c_{\mathrm{F}}(\widehat{s}).$$

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

$$m \leqslant c'_{\mathrm{F}}(m) s_0.$$

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 $m^* = \min\{m : m > c'_{\mathrm{F}}(m)s_0\}$ implies that

$$\widehat{s} < m^* + s_0$$

provided that the condition on m^* given by $t^{1/2} \geqslant 2\varphi_{\min}(m^* + s_0)(G)^{-1}\|\mathbb{E}_n[x_i\varepsilon_i]\|_\infty$ is met.

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 any 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 $(\mathrm{P} = \mathrm{P}_n)_{n\in\mathbb{N}}$. 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 $j \leqslant p$. The threshold satisfies a bound $t = O(\log p/n)$. In addition, $t^{1/2} \geqslant 2\varphi_{\min}(K_n)(G)^{-1}\|\mathbb{E}_n[x_i\varepsilon_i]\|_\infty$ with probability $1 - o(1)$.[5]

The rates assumed in Condition 1 reflect typical rates achieved under various possible sets of low-level conditions standard in the literature (ie. Belloni et al. (2012)). 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. In standard regression analysis where the number of covariates is small relative to the sample size, a common assumption used in establishing 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 $p \leqslant n$. 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 Bickel et al. (2009). This analysis focuses on a simple eigenvalue condition that was used in Belloni et al. (2012). Condition 1 could be shown to hold under more primitive conditions by adapting arguments found in Belloni and Chernozhukov (2013), which build upon results in Zhang and Huang (2008) and Rudelson and Vershynin (2008); see also Rudelson and Zhou (2013). Condition 1 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 $j \in S$, $\|\mathbb{E}_n[x_{iS}x_{iS}']^{-1}\mathbb{E}_n[x_{iS}x_{ik}]\|_1$ is strictly less than 1. The normalization $\mathbb{E}_n[x_{ij}^2] = 1$ is used to keep exposition concise and can be

[5] Formally, for a sequence of random variables X_n, the statement "$X_n = O(1)$ with probability $1 - o(1)$" is defined as: "there is a constant C independent of n such that $\mathrm{P}(|X_n| > C) \to 0$."

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 $\check{x}_{ij}$ denote x_{ij} residualized away from previously selected regressors and renormalized. Then even if $\mathbb{E}_n[x_{ij}\varepsilon_i] < t^{1/2}$, $\mathbb{E}_n[\check{x}_{ij}\varepsilon_i]$ 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 $\widehat{s}$ does not grow faster than s_0. Furthermore, this requirement does not implicitly impose a β-min condition and does not implicitly impose irrepresentability. The requirements on t can be relaxed if there is additional control on quantities of the form $\mathbb{E}_n[\check{x}_{ij}\varepsilon_i]$. Relative to analogous Lasso bounds in Belloni et al. (2012), Theorem 1 does not involve maximal sparse eigenvalues. This may become relevant if the components of x_i arise from factor model structures.

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 1 is considered in the next section, where analysis of 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 $\widehat{\theta}$ is obtained by Algorithm 1. Then there are bounds*

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

$$\widehat{s} \leqslant O(s_0),$$

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

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 $\mathbb{E}_n[(x_i'\theta_0 - x_i'\widehat{\theta})^2]^{1/2}$) 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 $\|\widehat{\theta} - \theta_0\|_1$ and $\|\widehat{\theta} - \theta_0\|_2$ of $\widehat{\theta}$ from the underlying unknown parameter θ_0. Theorem 3 above shows that deviations of $\widehat{\theta}$ 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 $\widehat{\theta}$ is obtained by Algorithm 1. Then there are bounds*

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

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

3. Testing-Based Forward Model Selection

The previous section presented results on convergence rates of Simple Forward Selection. The results of Theorem 1 are useful in 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 to the earlier one. Again, the observed data is given by $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, is generated by P, and $y_i = x_i'\theta_0 + \varepsilon_i$ 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 $\mathcal{E} : \mathbb{R}^p \to \mathbb{R}$ by

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

where E is expectation with respect to P. Note that $\mathcal{E}(\theta) = \mathrm{E}\ell(\theta)$. Extend the definition of $\mathcal{E}$ to apply also as a map $\mathcal{E} : 2^{\{1,\ldots,p\}} \to \mathbb{R}$ by $\mathcal{E}(S) = \min_{\theta:\mathrm{supp}(\theta)\subseteq S} \mathcal{E}(\theta)$. Similarly to before, for any S define the incremental loss from the jth covariate by

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

Within the class of greedy algorithms, it would be preferable to consider a greedy algorithm which inductively selects the jth covariate to enter a working model if $\Delta_j\mathcal{E}(S)$ is large in absolute value and $-\Delta_j\mathcal{E}(S) \geqslant -\Delta_k\mathcal{E}(S)$ 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 $T_{j\widehat{S}\alpha} = T_{k\widehat{S}\alpha} = 1$, the selection is made according to the larger value of $W_{j\widehat{S}}, W_{k\widehat{S}}$.

The use of $T_{jS\alpha}$ to define the set of covariates *eligible* for entry into the model, and W_{jS} to select *which* eligible covariate actually enters is conceptually important: it dissociates and highlights the two fundamental tasks of regularization and fitting.

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 $\widehat{S} = \varnothing$

For $1 \leqslant k \leqslant p$

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

3.2. Formal Analysis

This section formally states conditions on the hypothesis tests and 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 α, δ_{test}, c_{test}, c'_{test}, $c''_{\text{test}} > 0$ such that each of the following conditions hold.

1. The tests have power in the sense that

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

2. The tests control size in the sense that

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

3. The tests are continuous in the sense that

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

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 are considered simply tools for model selection, which coincidentally have many properties in common with traditional inferential hypothesis tests.

Condition 3 (*Regularity*). Normalizations $\mathrm{E}[\mathbb{E}_n[x_{ij}^2]] = 1$ hold for all j. The residuals decompose into $\varepsilon_i = \varepsilon_i^{\mathrm{o}} + \varepsilon_i^{\mathrm{a}}$ where $\mathrm{E}[\mathbb{E}_n[\varepsilon_i^{\mathrm{o}2}]] < \infty$, $\mathrm{E}[\mathbb{E}_n[\varepsilon_i^{\mathrm{o}} x_{ij}]] = 0$ for all j, and $\mathrm{E}[\mathbb{E}_n[\varepsilon_i^{\mathrm{a}2}]] \leqslant \frac{1}{2}\varphi_{\min}(K_{\text{test}})(\mathrm{E}[G])^{-1}c'_{\text{test}}$. Finally, $(80\times\varphi_{\min}(K_{\text{test}})(\mathrm{E}[G])^{-4}{c''_{\text{test}}}^{-3} + 1)s_0 < K_{\text{test}}$.

Condition 3 imposes regularity conditions for the class of models considered in the following theorem. First, ε_i is decomposed into an orthogonal component $\varepsilon_i^{\mathrm{o}}$ and an approximation component $\varepsilon_i^{\mathrm{a}}$, 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 $\mathrm{E}[G]$ 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 Belloni et al. (2012). The fact that $\varepsilon_i^{\mathrm{a}}$ 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 on the relative values of the sparse eigenvalues of $\mathrm{E}[G]$, c''_{test}, s_0, and K_{test}. Note that K_{test} measures the size of the set $S \subset \{1, ..., p\}$ over which the hypothesis tests 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 $\mathrm{E}[G]$ has small eigenvalues, when c''_{test} is small, or when s_0 is large.

There are a few cases where Condition 3 can be simplified. If $p > n$, even though the empirical Gram matrix is necessarily rank deficient, the population Gram matrix may be full rank. When $\mathrm{E}[G]$ is full rank, then $\lambda_{\min}(\mathrm{E}[G])$ may be used in place of $\varphi_{\min}(K_{\mathrm{test}})(\mathrm{E}[G])$. In addition, the condition on $\varepsilon_i^{\mathrm{a}}$ implicitly imposes constraints on c'_{test} and $\varphi_{\min}(K_{\mathrm{test}})(\mathrm{E}[G])^{-1}$. When there is no approximation error, this requirement is no longer needed.

Let

$$
\begin{aligned}
c_{\mathrm{T}} &= s_0 \varphi_{\min}(K_{\mathrm{test}})(\mathrm{E}[G])^{-1} c_{\mathrm{test}} \\
c'_{\mathrm{T}} &= 80 \times \varphi_{\min}(K_{\mathrm{test}})(\mathrm{E}[G])^{-4} {c''_{\mathrm{test}}}^{-3} \\
c''_{\mathrm{T}}(\widehat{s}) &= \varphi_{\max}(\widehat{s}+s_0)(G)^{1/2} \varphi_{\min}(\widehat{s}+s_0)(G)^{-1/2} \widehat{s}^{1/2} \left\| \mathbb{E}_n[x_i \varepsilon_i] \right\|_\infty \\
&\quad + 3 \varphi_{\max}(\widehat{s}+s_0)(G) \varphi_{\min}(\widehat{s}+s_0)(G)^{-1/2} (\widehat{s}+s_0)^{1/2} {c_{\mathrm{test}}}^{1/2} \varphi_{\min}(K_{\mathrm{test}})(\mathrm{E}[G])^{-1}.
\end{aligned}
$$

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

$$
\mathcal{E}(\widehat{S}) - \mathcal{E}(S_0) \leqslant c_{\mathrm{T}}
$$

$$
\widehat{s} \leqslant (c'_{\mathrm{T}} + 1)\, s_0
$$

$$
\mathbb{E}_n[(\, x_i'\theta_0 - x_i'\widehat{\theta}\,)^2]^{1/2} \leqslant c''_{\mathrm{T}}(\widehat{s})
$$

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

Theorem 4 provides finite sample bounds on the performance of TBFMS. In contrast to Theorem 1, 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 and estimation error attain the rate $O(s_0 \log p/n)$. This convergence rate matches typical Lasso and Post-Lasso rates.

The results aim to control the hybrid objectives, described in the introduction, of producing a good fit and returning a sparse set of variables. One useful implication of bounds controlling both $\widehat{s}$ and $\mathbb{E}_n[(x_i'\theta_0 - x_i\widehat{\theta})^2]$ is that the results can be applied to constructing uniformly valid post-model selection inference procedures (see Belloni et al. (2014)), in which for some applications, the prediction error bound alone is insufficient.

4. Examples and Extensions

This section describes an example application of Theorem 4. The main theoretical application is an illustration with heteroskedastic data. For this setting, a TBFMS procedure is constructed for which optimal convergence rates are proven.

4.1. 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. A TBFMS procedure is constructed based on the Heteroskedasticity-Consistent standard errors described in White (1980). The conditions required for the application of Theorem 4 are verified under low-level conditions on data generating processes. Other TBFMS procedures are possible, and these are discussed in the next section. The analysis begins with a formulation stated in Kozbur (2017b) (which does not derive nor claim any theoretical properties.)

For shorthand, write x_{ijS} (with $j \notin S$) 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}.

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

Define $\widehat{\varepsilon}_{ijS} = y_i - x_{ijS}'\widehat{\theta}_{jS}$. One heteroskedasticity robust estimate of the sampling variance of $\widehat{\theta}_{jS}$, proposed in White (1980), is given by the expression

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

where

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

Define the test statistics

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

Reject H_0 for large values of W_{jS}^{het} 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

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

The $\widehat{\tau}_{jS}$ will be helpful in addressing the fact that many different model selection paths are possible under different realizations of the data under P.[6] Not taking this fact 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 $W_{jS} = W_{jS}^{\mathsf{het}}$. Assign

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

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 $c_\tau \widehat{\tau}_{jS}$ 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 4 (*Regularity for Data with Heteroskedasticity*). Consider a sequence of data sets $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n \sim \mathrm{P} = \mathrm{P}_n$. 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 $\varepsilon_i = \varepsilon_i^{\mathrm{o}} + \varepsilon_i^{\mathrm{a}}$ such that a.s., $\mathrm{E}[\varepsilon_i^{\mathrm{o}}|x_i] = 0$ and $\max_i |\varepsilon_i^{\mathrm{a}}| = O(n^{-1/2})$. In addition, a.s., uniformly in i and n, $\mathrm{E}[\varepsilon_i^4|x_i]$ are bounded above and $\mathrm{E}[\varepsilon_i^2|x_i]$ is bounded away from zero. The covariates satisfy $\max_{j\leqslant p} \mathbb{E}_n[x_{ij}^{12}] = O(1)$ with probability $1 - o(1)$. There is a sequence K_n, where $s_0 = o(K_n)$, and bounds $\varphi_{\min}(K_n)(G)^{-1} = O(1)$, $\varphi_{\max}(K_n)(G) = O(1)$, $\varphi_{\min}(K_n)(\mathbb{E}_n[(\varepsilon_i x_i)(\varepsilon_i x_i)'])^{-1} = O(1)$, and $\max_{|S|<K_n, j\notin S} \|\eta_{jS}\|_1 = O(1)$, which hold with probability $1 - o(1)$. The rate condition $K_n^4 \log^3 p/n = o(1)$ holds.

Condition 4, as before, gives conditions on the sparse eigenvalues, this time applying to both G and to $\mathbb{E}_n[(\varepsilon_i x_i)(\varepsilon_i x_i')]$. In addition, Condition 4 assumes a bound on η_{jS} that may be strong in some cases. Previous results in Tropp (2004), Zhang (2009) 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 Meinshausen and Buhlmann (2006), use conditions analogous to $\max_{|S|\leqslant K_n, j\notin S} \|\eta_{jS}\|_1 = O(1)$ 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, because its proof does not leverage bounds relating $\widehat{W}_{jS}$ to the self-normalized sums $\mathbb{E}_n[x_{ij}'\varepsilon_i]/\sqrt{\mathbb{E}_n[x_{ij}^2\varepsilon_i^2]}$, $j \leqslant p$. Failure of the $O(1)$

[6] There is an unfortunate misprint in a Papers and Proceedings version of this paper, Kozbur (2017b), in which the exponent 1/2 is missing from the term $\mathrm{Diag}(\Psi_{jS}^{\widehat{\varepsilon}})$.

bound would lead only to slightly slower convergence rates. Condition 4 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 Jing et al. (2003)). Finally, the rate condition controls relative sizes of s_0, p, n because $s_0 < K_n$.

Theorem 5. *Consider a sequence of data sets $\mathcal{D}_n \sim \mathrm{P} = \mathrm{P}_n$ which satisfies Condition 4. Suppose that $c_\tau > 1$ is fixed independent of n, and that $\alpha = o(1)$ with $n\alpha \to \infty$.*[7] *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 $n \to \infty$.

4.2. Additional TBFMS Formulations and Variants

In general, the quality of statistical performance of a variant of TBFMS may depend on the structure of the data at hand through the size, power, and continuity properties of the tests, as articulated in Condition 2. Theorem 4 is general and can thus potentially be applied for many different types of TBFMS procedures, depending on how W_{jS} and $T_{jS\alpha}$ are defined in a particular setting. Depending on the setting, some variants may have better size, power, and continuity properties than others. This section describes several variants of the TBFMS procedure defined in the previous section.

The definition of the first variant considered is based on the observation 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 $\alpha > 0$ be parameters. Assign

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

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 and Stata. It is unknown to the author at the time of this writing whether the same convergence rates can be attained using the simpler tests under the identical regularity as in Condition 4. 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. The tests in Definition 2 are not necessarily more conservative than those in Definition 1.

The next variant TBFMS procedure for heteroskedastic data illustrates an important aspect of the result of Theorem 4. Namely, Theorem 4 explicitly allows the researcher to easily dissociate

[7] Allowing α to be fixed is possible under more restrictive conditions on the approximation error terms $\varepsilon_i^{\mathrm{a}}$. If $p > n$, then the rate $\log(p/\alpha)$ becomes equivalent to simply $\log(p)$.

the *regularization* component of high-dimensional estimation with the *fitting* component. For instance, the following formulation may be described as slightly greedier than Definition 2.

Definition 3 (*Fit-Streamlined Hypothesis Tests for Heteroskedastic Disturbances*)**.** Let $\alpha > 0$ be a parameter. Assign $W_{jS} = \Delta_j \ell(S)$. Assign

$$T_{jS\alpha} = 1 \iff W_{jS}^{\mathsf{het}} \geqslant \Phi^{-1}(1 - \alpha/p).$$

Under the more conservative tests, $T_{jS\alpha} = 1 \iff W_{jS}^{\mathsf{het}} \geqslant c_\tau \widehat{\tau}_{jS} \Phi^{-1}(1 - \alpha/p)$ for some $c_\tau > 1$, the same convergence rates of this greedier TBFMS procedure under Condition 4 are proven in the same way as Theorem 5 (by showing that the ratios $\Delta_j \ell(S)/W_{jS}^{\mathsf{het}}$ are sufficiently well behaved). For brevity, the details are omitted.

When the data is approximately homoskedastic, the tests defined in Definition 1 may be too conservative and suffer in terms of power (noting that power is an explicit input into the bounds in Theorem 4). In this case, tests of the following form can be considered using the homoskedastic-based test statistics $W_{jS}^{\mathsf{hom}} = [\widehat{V}_{jS}^{\mathsf{hom}}]_{jj}^{-1/2} \left| [\widehat{\theta}_{jS}]_j \right|$ with $\widehat{V}_{jS}^{\mathsf{hom}} = \frac{1}{n} \mathbb{E}_n[\widehat{\varepsilon}_{ijS}^2] \mathbb{E}_n[x_{ijS} x_{ijS}']^{-1}$.

Definition 4 (*Simplified Hypothesis Tests for Homoskedastic Disturbances*)**.** Let $\alpha > 0$ be parameters. Assign $W_{jS} = W_{jS}^{\mathsf{hom}}$. Assign

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

Evidence from the simulation study suggests that this option performs better than the more complex tests in Definition 1 when the data is homoskedastic, but not when heteroskedastic.

4.3. TBFMS with Baseline Covariates

The initialization statement of Algorithm 2 can be modified trivially so that a baseline set S_{base} of covariates are automatically included in $\widehat{S}$ at the start of selection. In this case, the initialization statement of Algorithm 2 is replaced with

Initialize. Set $\widehat{S} = S_{\text{base}}$.

Under this modification, a direct analogue of Theorem 4 holds. It is proven using the same arguments. Sparsity bounds can be calculated as in the proof of Theorem 4 by separately tracking covariates $S_0 \setminus S_{\text{base}}$. This requires an appropriate adjustment to Condition 3 in which K_{test} must be larger, and in particular, bound a quantity depending on both s_0 and $|S_{\text{base}}|$. The proof of the first and third statements of Theorem 4 do not depend on the initialization $S = \varnothing$. Such modification is appropriate, for instance, in cases in which a researcher wishes to include a set of covariates into a model, but is unsure of which interactions to include. In this case, TBFMS can be used to help identify relevant interaction terms. This case is further explored in the empirical application. Similarly, a constant term may be included automatically in the regression model.

4.4. Additional Discussion of Potential Variants

Analogous results potentially hold for dependent data using HAC-type standard errors (see Newey and West (1987), Andrews (1991).) In addition, cluster-type standard errors for large-T-large-n and fixed-T-large-n panels can be used by adapting arguments from Belloni et al. (2016). Analogous results for homoskedastic disturbances can be derived as a corollary.

Another alternative 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 $H_0 : \Delta_j \mathcal{E}(S) = 0$ for $j \leqslant p$ 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.

5. Example: TBFMS for Asset-Based Poverty Mapping

This section investigates the use of TBFMS to develop improved proxy-means tests in an application to poverty mapping in a Peruvian dataset covering years 2010–2011. This analysis extends the original analysis in Hanna and Olken (2018), who estimated a predictive model of household consumption using the same data. The data is from the Peruvian Encuesta Nacional de Hogares (ENAHO), maintained by the Instituto Nacional de Estadistica e Informatica (INEI), Peru.

Fighting poverty is a major priority for many developing countries and international organizations like the United Nations. Strategies for combating poverty often require that governing bodies have accurate information about household level consumption, income, or other measures of welfare. Methods for empirical identification of households and regions below a given poverty line is are considered in e.g. Elbers et al. (2003), among others.

One method in which a government can obtain a signal about measures of welfare is termed a *proxy-means test.* The implementation of a proxy-means test is usually based on large censuses of the population, in which government enumerators obtain information on easily observable and verifiable assets. The government uses these assets to predict incomes or per-capita consumption or other measures of poverty or welfare by estimating a regression on a smaller sample with detailed measurement of consumption. The *proxy-means score* is defined as the predicted income or consumption, which is calculated using the results from the predictive regression. This method is widespread, and is implemented in several countries including Indonesia, Pakistan, Nigeria, Mexico, Philippines, Burkina Faso, Ecuador, Jamaica. Improved predictions may be helpful to policy makers in deciding on strategies to eliminate poverty; see Fiszbein et al. (2009).

The data contain covariates which are indicator variables describing household-level assets and which are used to predict outcomes $y =$ Consumption (in 10^3 Peruvian Soles) as well as $y =$ *log* Consumption (in Peruvian Soles). The set of 46305 household observations is split randomly (with equal probability) into a training sample of size 22674 and a testing sample of size 22704. All estimation procedures considered are implemented on the training sample. The indicators

derive from factor variables describing a household's (1) water source, (2) drain infrastructure (3) wall material, (4) roof material, (5) floor material, (6) availability of electricity, (7) access to telephone, (8) education of head of household, (9) type of insurance, (10) crowding, (11) consumption of luxury items. See Hanna and Olken (2018) for more details.

Here, TBFMS is used to determine which interactions of underlying covariates described above are helpful in developing an improved proxy-means test.[8] For every (unordered) pair of indicator variables A and B, three symmetric logic functions,

$$\mathsf{and}(A,B),\ \mathsf{or}(A,B),\ \mathsf{xor}(A,B).$$

are generated. Together with a constant term, these logic functions linearly span exactly the set of symmetric boolean functions on all pairs of indicators. In addition, for every unordered triple of indicator variables, A, B and C, the function

$$\mathsf{and}(A,B,C)$$

is generated. The final interaction expansion is based on the union of the generated functions.

Proxy-means tests are estimated with ordinary least squares (OLS) using un-interacted indicator variables (replicating Hanna and Olken (2018)) and with four TBFMS estimators adding interactions. The TBFMS estimators are *TBFMS I–IV* and use Algorithm 2 together with tests of Definitions 1–4, with selection initialized, as described in Section 4.3, with baseline variables S_{base} consisting of original (un-interacted) indicators and a constant term. Tuning parameters are set to $c_\tau = 1.01, \alpha = .05$ in all cases. Including baseline indicators is natural because they are ex-ante researcher chosen as relevant and because they are few in number relative to the sample size. 10 baseline indicators of the 72 contained in the original data were excluded due to multicollinearity (after having generated interactions), leaving 62 baseline indicators (which includes a constant term). Generated interaction terms with training sample standard deviation < 0.03 (corresponding to 20 or less 1s) were discarded. Interactions are residualized away from baseline indicators on the training sample[9]. Residualizing only affects the terms $\widehat{\tau}_{jS}$ in *TBFMS I*, with $\widehat{\tau}_{jS}$ typically becoming smaller, thus resulting in more selections into the model. Residualizing does not affect the estimation *TBFMS II—IV*. The final number of generated interactions is 23964 which exceeds the training sample size of 22674.

Table 1 reports test sample mean square error (MSE), training sample MSE, relative and absolute reduction in test sample MSE (compared to OLS), p-values and confidence sets for absolute reduction in MSE based two-sided paired t-tests. Table 1 also reports summary statistics.

[8] Hanna and Olken (2018) also note that other estimation techniques may offer improvement. The reference Nichols and McBride (2018) also investigates various cross validation and machine learning techniques toward this end.

[9] Residualizing does not involve the outcome variable and can be viewed similarly to the process of demeaning covariates provided the number of baseline indicators is sufficiently small. Residualizing introduces a small amount of dependence across observations as does demeaning. The analysis in the proof of Theorem 5 can be adjusted to show that dependence arising from residualization away from a fixed, suitably small set of covariates has negligible effects.

TABLE 1
Asset-Based Poverty Estimation

	# Var	Train MSE	Test MSE	Rel(%) ΔMSE	ΔMSE	*p*-val ΔMSE	95% CI ΔMSE	
	A. y = Consumption ($\times 10^3$ Peruvian Soles)							
Estimates								
OLS	62	0.0780	0.0897	–	–	–	–	–
TBFMS I	116	0.0727	0.0867	3.3503	0.0030	2.3E-08	(0.0020	0.0041)
TBFMS II	159	0.0697	0.0860	4.1264	0.0037	0.0002	(0.0018	0.0056)
TBFMS III	118	0.0654	0.0848	5.4865	0.0049	0.0002	(0.0023	0.0076)
TBFMS IV	118	0.0654	0.0848	5.4865	0.0049	0.0002	(0.0023	0.0076)
	B. y = *log* Consumption (Peruvian Soles)							
Estimates								
OLS	62	0.1916	0.1910	–	–	–	–	–
TBFMS I	117	0.1849	0.1884	1.3954	0.0027	2.1E-07	(0.0017	0.0037)
TBFMS II	140	0.1829	0.1892	0.9669	0.0018	0.0109	(0.0004	0.0033)
TBFMS III	89	0.1818	0.1865	2.3809	0.0045	1.8E-10	(0.0031	0.0060)
TBFMS IV	89	0.1818	0.1865	2.3809	0.0045	1.8E-10	(0.0031	0.0060)

Data description	Training sample size : 22674, Test sample size : 22704						
	# Baseline characteristics : 62, # Interacted characteristics : 23964						
	Years collected : 2010–2011						
	Source : Instituto Nacional de Estadistica e Infomatica (INEI)						
	Encuesta Nacional de Hogares (ENAHO); also used in Hanna and Olken (2018)						
Outcome distribution summary (unconditional)							
		Q05	Q50	Q95	Mean	Stdev	Var
y = Consumption ($\times 10^3$ Peruvian Soles)		0.1015	0.3550	1.1582	0.4608	0.4136	0.1711
y = *log* Consumption (Peruvian Soles)		4.6254	5.8766	7.0516	5.8566	0.7442	0.5539

Asset-based proxy-means estimation results. Estimates are presented for *TBFMS I–TBFMS IV*, *OLS* as described in the text. The *OLS* estimator replicates analysis in Hanna and Olken (2018).

Each TBFMS estimate indicates significant improvement in test sample MSE relative to OLS at the 5% level. Highest relative improvement are with *TBFMS III* and *TBFMS IV*, which selected the same models both for y = *log* Consumption and y = Consumption. For y = Consumption, in the test sample, *TBFMS III, IV* give a 5.4865% relative improvement to *OLS* in MSE; 0.0049 absolute improvement, p-value = .0002 based on 2-sided paired t-test, 95% CI (0.0023, 0.0076). For *log* Consumption, *TBFMS III, IV* give a 2.3809% relative improvement to OLS in MSE, absolute 0.0045 improvement; p-value = 1.8E-10 based on 2-sided paired t-test, 95% CI (0.0031, 0.0060).

6. *TBFMS* Simulation Studies

The results in the previous sections suggest that estimation with TBFMS should produce quality estimates in large sample sizes. This section conducts two simulation studies to evaluate the finite sample performance of TBFMS relative to select other procedures commonly used in high-dimensional regression problems in finite samples.

The simulation study compares the following estimators.

1. *TBFMS I.* Algorithm 2 with tests defined in Definition 1 with $c_\tau = 1.01$, $\alpha = .05$.
2. *TBFMS II.* Algorithm 2 with simplified tests defined in Definition 2 with $\alpha = .05$.
3. *TBFMS III.* Algorithm 2 with streamlined tests defined in Definition 3 with $\alpha = .05$.
4. *TBFMS IV.* Algorithm 2 with homoskedastic tests defined in Definition 4 with $\alpha = .05$.
5. *Lasso-CV.* Standard Lasso, with penalty parameter chosen using 10-fold cross validation. $\widehat{\theta}$ is the minimizer of the Lasso objective function and is not refit on the selected model.
6. *Post-Het-Lasso.* Lasso implementation in Belloni et al. (2012), which is designed specifically to handle heteroskedastic disturbances. This requires two tuning parameters to c_τ and α set to $c_\tau = 1.01$ and $\alpha = .05$. The estimate $\widehat{\theta}$ is refit on the selected model.
7. *Oracle* Least squares regression on model consisting of $S_0 = \{j : [\theta_0]_j \neq 0\}$.

6.1. Simulation I

The first simulation study, Simulation I, evaluates TBFMS relative to several alternative estimators on a series of performance metrics for high-dimensional sparse linear regression problems. Simulation I considers data of the form $\mathcal{D}_n = \{(y_i, x_i)\}_{i=1}^n$ with $y_i = x_i'\theta_0 + \varepsilon_i$. Samples $\mathcal{D}_n$ are drawn from several data-generating processes reported in Table 2. Each considered data generating process is characterized by parameters s_0, ρ_0, b_0, n. The parameter s_0 is the sparsity. The parameter b_0 controls the nature of the coefficient vector by $\theta_{0j} = b_0^{j-1}\mathbf{1}_{j \leqslant s_0}$. When $b_0 < 0$, the coefficients θ_{0j} alternate sign in j and when $|b_0| < 1$, the coefficients decay. 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. The dimensionality is always taken to be double the sample size so that $p = 2n$. Each simulation design is replicated 1000 times.

The simulation results for Simulation I are reported in Table 3. The results track various measures of estimation quality for the five estimators for fixed s_0, b_0, and ρ_0, and for n. Columns in Table 3 display (1) *MPEN*—mean prediction error norm defined as $\mathbb{E}_n[(x_i'\theta_0 - x_i'\widehat{\theta})^2]^{1/2}$, (2) *RMSE*—root mean square estimation error defined as $\|\widehat{\theta}_2 - \theta_0\|_2$, (3) *MNCS*—mean number of correctly selected covariates from S_0, (4) *MSSS*—mean size of selected set of covariates in all cases averaged over simulation replications.

Table 3 indicates that predictive and estimation performance improve for all estimates as n increase from 100 to 500. *Lasso-CV* selects an increasing number of variables. Outside the *Oracle* estimator, for all n, *TBFMS II* attains the best predictive and estimation performance in settings with heteroskedasticity. In Table 3, 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 - IV* and *Post-Het-Lasso.* With alternating coefficients ($b_0 = -0.5$), *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 θ_0 relative to the correlation structure of the covariates. Finally, the relative difference in performance across estimators is larger in the presence of heteroskedas-

ticity. 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 Chetverikov et al. (2016) for analysis of Lasso with cross-validation). But it is seen from this simulation study that *Lasso-CV* leads to selection of substantially more covariates selected.

6.2. Simulation II

This section performs a second simulation study, Simulation II, on the use of TBFMS in selecting control covariates for the estimation of the effect of a covariate of interest on an outcome from a large set of potential observable controls.

Simulation II considers data of the form $\mathcal{D}_n = \{(y_i, x_i, w_i)\}_{i=1}^n \sim \mathrm{P}_n$ 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. In particular,

$$y_i = x_i\beta_0 + w_i'\theta_0^1 + \varepsilon_i$$
$$x_i = w_i'\theta_0^2 + u_i$$

for some parameters $\beta_0 \in \mathbb{R}, \theta_0^1, \theta_0^2 \in \mathbb{R}^p$ with $\mathrm{E}[\varepsilon_i|w_i, x_i] = 0$ and $\mathrm{E}[u_i|w_i] = 0$. Here, the impact of the policy/treatment variable x_i on the outcome y_i is measured by the unknown parameter β_0 which is the target of inference. The w_i are potentially important conditioning variables. The confounding factors w_i affect x_i via the function $w_i'\theta_0^2$ and the outcome variable via the function $w_i'\theta_0^1$. Both of the parameters θ_0^1 and θ_0^2 are unknown. As in Simulation I, specific data-generating processes used in Simulation II depend on parameters s_0, b_0, ρ, n and are given by Table 2.

The structure of the TBFMS procedure applied to a specific linear regression problem ensures that any coefficient in that problem, unless reliably distinguishable from zero, is estimated to be exactly zero. This property complicates inference after model selection in sparse models that may have a set of covariates with small but non-zero coefficients. The use of TBFMS for a linear regression y_i on (x_i, z_i), possibly initializing the $\widehat{S}$ to including x_i at the start of model selection, may result in excluding important conditioning covariates, which may lead to non-negligible omitted variables bias of parameters of interest. As a result, inference which does not take account the possibility of such model selection mistakes can be distorted. This intuition is formally developed in Leeb and Pötscher (2008). Several recent papers offer solutions to this problem; see, for example, Zhang and Zhang (2014), Belloni et al. (2014), van de Geer et al. (2014); Javanmard and Montanari (2014).

To estimate β_0 in this environment, adopt the *post-double-selection* method of Belloni et al. (2014) in conjunction with TBFMS. 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:

$$y_i = w_i'\theta_0^{\mathrm{RF}} + v_i$$
$$x_i = w_i'\theta_0^{\mathrm{FS}} + u_i$$

with $\theta_0^{\text{FS}} = \theta_0^2$, referred to as the *first stage* (FS) coefficient, and $\theta_0^{\text{RF}} = \theta_0^1 + \beta_0\theta_0^2$, referred to as the *reduced form* (RF) coefficient. 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. Belloni et al. (2014) develop and discuss the post-double-selection method in detail. Using two model selection steps guards against distorted inference and guarantees that variables excluded in both model selection steps have a negligible contribution to omitted variables bias.

Analogously to Simulation Study I, Seven estimators are considered for estimation and are named *TBFMS I –TBFMS IV, Lasso-CV, Post-Het-Lasso*, and *Oracle*. 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 above. Final variance estimates $\widehat{V}$ for $\widehat{\beta}$ are based on HC3 standard errors White (1980). For each estimator and simulation setting, the bias and standard deviation of the point estimates, and coverage probability and average interval length are computed over 1000 simulation replications. Results are shown in Table 4.

The simulation results indicate that across the data-generating processes considered, TBFMS estimators largely achieve near-*Oracle* coverage probabilities. *TBFMS IV*-based estimates exhibit some size distortion in heteroskedastic settings. In all settings, bias, standard deviation, and interval lengths of *TBFMS I – III* closely resemble the *Oracle* estimator. In some simulations (notably in Panel A) there is a large difference in coverage probabilities between the TBFMS estimates and the *Post-Het-Lasso* estimate. Though the *Post-Het-Lasso*-based confidence sets are asymptotically uniformly valid, and are theoretically robust against model selection mistakes, finite sample model selection properties remain important. Interestingly, in this case, using the relaxed penalty level with *Lasso-CV* does not improve coverage probability.

TABLE 2
Design for Simulation Studies I and II

I. High-Dimensional Prediction.		**II. High-Dimensional Controls**.	
Data:	$\mathcal{D}_n = (y_i, x_i)_{i=1}^n$ *iid*	Data:	$\mathcal{D}_n = (y_i, x_i, w_i)_{i=1}^n$ *iid*
DGP:	$y_i = x_i\theta_0 + \varepsilon_i$	DGP:	$y_i = x_i\beta_0 + w_i'\theta_0^1 + \varepsilon_i, \quad x_i = w_i'\theta_0^2 + u_i$
	$p = \dim(x_i) = 2n$		$p = \dim(w_i) = 2n$
	$\theta_{0j} = b_0^{j-1}\mathbf{1}_{j\leqslant s_0}$		$\theta_{0j}^1 = b_0^{j-1}\mathbf{1}_{j\leqslant s_0}, \ \theta_{0j}^2 = \sin(j)\mathbf{1}_{j\leqslant s_0}$
	$x_{ij} \sim N(0,1), \ \text{corr}(x_{ij}, x_{ik}) = 0.5^{\lvert j-k\rvert}$		$w_{ij} \sim \text{N}(0,1), \ \text{corr}(w_{ij}, w_{ik}) = 0.5^{\lvert j-k\rvert}$
	$\varepsilon_i \sim \sigma_i N(0,1)$		$(\varepsilon_i, u_i) \sim \sigma_i \text{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
	$\sigma_i = \exp(\rho_0 \sum_{j=1}^p 0.75^{(p-j)} x_{ij})$		$\sigma_i = \exp(\rho_0 \sum_{j=1}^p 0.75^{(p-j)} w_{ij})$
Settings:	$s_0 = 6$, $b_0 \in \{-0.5, 0.5\}$, $\rho_0 \in \{0, 0.5\}$	Settings:	$s_0 = 6$, $b_0 \in \{-0.5, 0.5\}$, $\rho_0 \in \{0, 0.5\}$
	$n \in \{100, 500\}$		$n \in \{100, 500\}$

TABLE 3
Simulation I Results: Prediction in the Linear Model

	$n = 100$				$n = 500$			
	MPEN	RMSE	MNCS	MSSS	MPEN	RMSE	MNCS	MSSS
A. $\rho_0 = 0$: Homoskedastic, $s_0 = 6$: High Sparsity, $b_0 = -0.5$: Alternating Coefficients								
TBFMS I	0.313	0.484	1.470	1.470	0.146	0.218	2.599	2.599
TBFMS II	0.192	0.281	2.330	2.444	0.094	0.132	3.403	3.478
TBFMS III	0.194	0.286	2.307	2.307	0.094	0.133	3.432	3.432
TBFMS IV	0.191	0.281	2.301	2.358	0.093	0.132	3.400	3.461
Post-Lasso	0.417	0.640	0.958	0.958	0.401	0.617	1.000	1.000
Lasso-CV	0.285	0.468	2.850	21.463	0.173	0.275	3.693	40.617
Oracle	0.117	0.149	6.000	6.000	0.053	0.066	6.000	6.000
B. $\rho_0 = 0.5$ Heteroskedastic, $s_0 = 6$: High Sparsity, $b_0 = -0.5$: Alternating Coefficients								
TBFMS I	0.507	0.720	0.815	0.815	0.395	0.593	1.077	1.077
TBFMS II	0.478	0.671	0.994	1.012	0.276	0.401	1.765	1.780
TBFMS III	0.526	0.734	0.824	0.824	0.294	0.429	1.650	1.650
TBFMS IV	0.529	0.726	0.896	0.973	0.299	0.425	1.713	1.807
Post-Het-Lasso	0.884	1.147	0.015	0.015	0.581	0.846	0.867	0.867
Lasso-CV	0.594	0.849	1.242	8.863	0.438	0.661	1.670	18.326
Oracle	0.379	0.482	6.000	6.000	0.180	0.222	6.000	6.000
C. $\rho_0 = 0$: Homoskedastic, $s_0 = 6$: High Sparsity, $b_0 = 0.5$: Positive Coefficients								
TBFMS I	0.307	0.395	2.331	2.331	0.147	0.186	3.396	3.396
TBFMS II	0.193	0.244	3.094	3.193	0.091	0.113	4.078	4.152
TBFMS III	0.193	0.244	3.064	3.064	0.091	0.114	4.105	4.105
TBFMS IV	0.191	0.241	3.062	3.109	0.091	0.114	4.068	4.128
Post-Het-Lasso	0.782	0.583	1.174	1.174	0.615	0.468	2.121	2.121
Lasso-CV	0.207	0.204	4.570	15.392	0.109	0.099	5.257	20.100
Oracle	0.117	0.149	6.000	6.000	0.053	0.066	6.000	6.000
D. $\rho_0 = 0.5$: Heteroskedastic, $s_0 = 6$: High Sparsity, $b_0 = 0.5$: Positive Coefficients								
TBFMS I	0.665	0.789	1.274	1.274	0.405	0.499	1.976	1.976
TBFMS II	0.513	0.617	1.759	1.780	0.285	0.346	2.580	2.590
TBFMS III	0.577	0.662	1.505	1.505	0.310	0.369	2.400	2.400
TBFMS IV	0.580	0.670	1.574	1.656	0.314	0.373	2.478	2.574
Post-Het-Lasso	1.314	1.014	0.288	0.288	0.879	0.663	1.898	1.898
Lasso-CV	0.570	0.522	3.125	11.661	0.334	0.289	3.873	16.337
Oracle	0.379	0.482	6.000	6.000	0.180	0.222	6.000	6.000

Simulation results for estimation in 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 estimators, *TBFMS I– TBFMS IV, Post-Het-Lasso, Lasso-CV*, and *Oracle* described in the text. Columns display (1) *MPEN*—mean prediction error norm defined as $\mathbb{E}_n[(x_i'\theta_0 - x_i'\widehat{\theta})^2]^{1/2}$, (2) *RMSE*—root mean square estimation error defined as $\|\widehat{\theta}_2 - \theta_0\|_2$, (3) *MNCS*—mean number of correctly selected covariates from S_0, (4) *MSSS*—mean size of selected set of covariates in all cases averaged over simulation replications. Figures are based on 1000 simulation replications.

TABLE 4
Simulation II Results: Control Selection in the Linear Model

	$n = 100$				$n = 500$			
	Bias	StDev	Length	Cover	Bias	StDev	Length	Cover
	D. $\rho_0 = 0$: Homoskedastic, $s_0 = 6$: High Sparsity, $b_0 = -0.5$: Alternating Coefficients							
TBFMS I	-0.035	0.129	0.354	0.772	-0.036	0.060	0.182	0.797
TBFMS II	-0.009	0.114	0.428	0.934	0.000	0.047	0.178	0.941
TBFMS III	-0.016	0.112	0.424	0.928	-0.000	0.047	0.178	0.941
TBFMS IIV	-0.011	0.112	0.425	0.929	-0.000	0.047	0.178	0.940
Post-Het-Lasso	-0.190	0.057	0.210	0.068	-0.193	0.025	0.094	0.000
Lasso CV	-0.193	0.054	0.214	0.067	-0.193	0.025	0.094	0.000
Oracle	0.002	0.105	0.424	0.956	-0.001	0.047	0.178	0.941
	B. $\rho_0 = 0.5$: Heteroskedastic, $s_0 = 6$: High Sparsity, $b_0 = -0.5$: Alternating Coefficients							
TBFMS I	0.032	0.326	1.016	0.881	0.003	0.219	0.752	0.927
TBFMS II	0.010	0.373	1.179	0.901	0.005	0.231	0.806	0.929
TBFMS III	0.010	0.366	1.141	0.897	0.004	0.231	0.801	0.928
TBFMS IV	0.009	0.372	1.167	0.899	0.005	0.231	0.803	0.929
Post-Het-Lasso	-0.111	0.183	0.553	0.721	-0.104	0.129	0.421	0.650
Lasso CV	-0.116	0.209	0.637	0.745	-0.104	0.130	0.422	0.651
Oracle	0.004	0.406	1.333	0.909	0.009	0.234	0.820	0.934
	C. $\rho_0 = 0$: Homoskedastic, $s_0 = 6$: High Sparsity, $b_0 = 0.5$: Positive Coefficients							
TBFMS I	0.012	0.122	0.362	0.814	-0.054	0.059	0.184	0.729
TBFMS II	-0.012	0.113	0.426	0.929	0.000	0.044	0.179	0.960
TBFMS III	-0.019	0.112	0.423	0.926	-0.000	0.044	0.179	0.961
TBFMS IV	-0.014	0.113	0.423	0.920	0.000	0.044	0.179	0.960
Post-Het-Lasso	-0.088	0.061	0.234	0.669	-0.085	0.027	0.102	0.109
Lasso CV	-0.086	0.059	0.234	0.680	-0.085	0.027	0.102	0.109
Oracle	0.002	0.102	0.420	0.954	-0.001	0.044	0.179	0.960
	D. $\rho_0 = 0.5$: Heteroskedastic, $s_0 = 6$: High Sparsity, $b_0 = 0.5$: Positive Coefficients							
TBFMS I	-0.032	0.334	1.020	0.857	0.009	0.250	0.763	0.918
TBFMS II	-0.016	0.371	1.196	0.882	-0.001	0.259	0.811	0.923
TBFMS III	-0.010	0.362	1.153	0.884	-0.001	0.257	0.807	0.920
TBFMS IV	-0.007	0.363	1.176	0.885	0.000	0.259	0.809	0.920
Post-Het-Lasso	-0.161	0.216	0.596	0.638	-0.044	0.154	0.427	0.859
Lasso CV	-0.063	0.203	0.641	0.903	-0.044	0.154	0.427	0.858
Oracle	-0.019	0.397	1.352	0.895	0.003	0.263	0.826	0.924

Simulation results for estimation in 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 estimators, *TBFMS I– TBFMS IV, Post-Het-Lasso, Lasso-CV*, and *Oracle* described in the text. Columns display (1)*Bias*—bias of the respective estimates for β_0, (2) 'textitStDev—standard deviation of the respective estimates for β_0, (3) *Length*—length of confidence intervals for β_0, (4) *Cover*—coverage probabilities of the respective 95% confidence intervals for β_0. Figures are based on 1000 simulation replications.

7. 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.

Appendix A: Proofs

This appendix proves Theorems 1 and 4. As TBFMS is a greedy procedure which is not the resulting solution of a simple optimization problem, the proofs establishing the properties of TBFMS cannot refer to any global optimality conditions. This fact limits the applicability of common m-estimation arguments or arguments for similar bounds for Lasso, and requires the development of certain different techniques.

In the course of the proofs, several important results along the way are recorded as lemmas. Lemmas which do not follow immediately from arguments in this section are proven in the online supplemental material, Supplement to "Analysis of Testing-Based Forward Model Selection."

In addition, Theorems 2, 3 and 5 are proven in the online supplemental material.

A.1. Proof of the First Statement of Theorem 1

The first statement of Theorem 1 is proven by creating an appropriate analogue of the *basic inequality*[10] from standard Lasso analysis. Specifically, the following lemma holds.

Lemma 1. $\ell(\widehat{S} \cup S_0) \leqslant \ell(\theta_0)$.

Lemma 1 holds by $\ell(\widehat{S} \cup S_0) \leqslant \ell(S_0) \leqslant \ell(\theta_0)$. Once the analogue basic inequality is noted, $\ell(\widehat{\theta})$ can be related to $\ell(\theta_0)$ with a bound that depends on s_0, t, and $\varphi_{\min}(\widehat{s}+s_0)(G)^{-1}$. The following is Lemma 3.3 in Das and Kempe (2011).

Lemma 2 (Das and Kempe (2011))**.** $\ell(\widehat{S}) - \ell(\widehat{S} \cup S_0) \leqslant \varphi_{\min}(\widehat{s}+s_0)(G)^{-1} \sum_{j \in S_0 \setminus \widehat{S}} (-\Delta_j \ell(\widehat{S}))$.

Using the fact that t is the threshold in Algorithm 1, and thus $-\Delta_j \ell(\widehat{S}) \leqslant t$, gives the further bound $\ell(\widehat{S}) - \ell(\widehat{S} \cup S_0) \leqslant s_0 t \varphi_{\min}(\widehat{s}+s_0)(G)^{-1}$. Applying the basic inequality along with the fact that $\ell(\widehat{S}) = \ell(\widehat{\theta})$ gives $\ell(\widehat{\theta}) \leqslant \ell(\theta_0) + s_0 t \varphi_{\min}(\widehat{s}+s_0)(G)^{-1}$. Expanding the quadratics $\ell(\widehat{\theta})$ and $\ell(\theta_0)$, and applying arguments analogous to those in Lemma 6 in Belloni et al. (2012), gives

Lemma 3. $\mathbb{E}_n[(x_i'\theta_0 - x_i'\widehat{\theta})^2]^{1/2} \leqslant c_{\mathrm{F}}(\widehat{s})$.

Lemma 3 is the first statement of Theorem 1.

A.2. Proof of First and Third Statements of Theorem 4

Let $\mathcal{T}$ be the event implied by Condition 2. Then $\mathrm{P}(\mathcal{T}) \geqslant 1 - \alpha - 3\delta_{\text{test}}/3 = 1 - \alpha - \delta_{\text{test}}$. The rest of the proof works on the event $\mathcal{T}$. If Algorithm 2 terminates at K_{test} steps or ear-

[10] For Lasso estimation with penalty level λ, the basic inequality asserts that $\ell(\widehat{\theta}) + \lambda\|\widehat{\theta}\|_1 \leqslant \ell(\theta_0) + \lambda\|\theta_0\|_1$.

lier, then it terminates at a step with $-\Delta_j \mathcal{E}(\widehat{S}) \leqslant c_{\text{test}}$ for every $j \notin \widehat{S}$. Similarly to the proof of Theorem 1 above, Lemma 3.3 in Das and Kempe (2011) yields $|\mathcal{E}(S_0) - \mathcal{E}(\widehat{S})| \leqslant \varphi_{\min}(K_{\text{test}})(\mathrm{E}[G])^{-1} \sum_{j \in S_0 \setminus \widehat{S}} -\Delta_j \mathcal{E}(S) \leqslant s_0 c_{\text{test}} \varphi_{\min}(K_{\text{test}})(\mathrm{E}[G])^{-1}$. It is shown in the next section that $\widehat{s} \leqslant K_{\text{test}}$ on $\mathcal{T}$, completing the proof of the first statement of Theorem 4. Next,

Lemma 4. $\mathbb{E}_n[(x_i'\theta_0 - x_i'\widehat{\theta})^2]^{1/2} \leqslant c_{\mathrm{T}}''(\widehat{s})$.

Lemma 4 is the third statement of Theorem 4.

A.3. Proof of Sparsity Bounds for Theorems 1 and 4

The sparsity bounds in Theorems 1 and 4 are proven together. In the case of Theorem 1, the covariates $x_j = (x_{1j}, ..., x_{nj})'$, outcome $y = (y_1, ..., y_n)'$ as well as disturbances $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$ are considered elements of the Hilbert space $\mathbb{R}^n$ with inner product $\langle a, b \rangle = n \times \mathbb{E}_n[a_i b_i]$. In the case of Theorem 4, x_j, y, ε are elements of the Hilbert space $L^2(\Omega, \mathbb{R}^n)$, of P-square-integrable random vectors taking values in $\mathbb{R}^n$ with Ω an underlying probability space, with inner product $\langle a, b \rangle_{L^2(\Omega, \mathbb{R}^n)} = n \times \mathrm{E}\mathbb{E}_n[a_i b_i]$. The notation H is used to denote the appropriate of these Hilbert spaces according to the cases of $\mathsf{H} = \mathbb{R}^n$ for Theorem 1 and $\mathsf{H} = L^2(\Omega, \mathbb{R}^n)$ for Theorem 4. In addition, let $t_{\mathsf{H}} = t$ and $G_{\mathsf{H}} = G$ in the case of Theorem 1 and let $t_{\mathsf{H}} = c'_{\text{test}}$ and $G_{\mathsf{H}} = \mathrm{E}[G]$ in the case of Theorem 4. In the case of Theorem 4, the arguments that follow hold on $\mathcal{T}$, the event defined by Condition 2 (see also the previous subsection).

A.3.1. Two Orthogonalizations

Let $v_k \in \mathsf{H}$, $k = 1, ..., s_0$ denote *true covariates* which refer to x_j for $j \in S_0$. The term *false covariates* refers to those x_j for which $j \notin S_0$. Consider the step after which there are exactly m false covariates selected into the model. These are denoted $w_1, ..., w_m$, ordered according to the order they were selected, and indexed by the set $A = \{1, ..., m\}$.

Apply Gram-Schmidt orthogonalization to $v_1, ..., v_{s_0}, \varepsilon$ with the inner product from H. The ordering is according to selection into $\widehat{S}$. Any true covariates unselected at m false covariate selections are temporarily ordered arbitrarily after the selected true covariates. ε is placed last. This yields a new set of unit-norm elements

$$v_1, ..., v_{s_0}, \varepsilon \mapsto \tilde{v}_1, ..., \tilde{v}_{s_0}, \tilde{\varepsilon} \in \mathsf{H}.$$

This orthogonalization also yields parameters

$$\tilde{\theta} = (\tilde{\theta}_1, ...\tilde{\theta}_{s_0})' \in \mathbb{R}^{s_0}, \ \tilde{\theta}_{\tilde{\varepsilon}} \in \mathbb{R} \ \text{ satisfying } \ y = \tilde{v}_1\tilde{\theta}_1 + ... + \tilde{v}_{s_0}\tilde{\theta}_{s_0} + \tilde{\theta}_{\tilde{\varepsilon}}\tilde{\varepsilon}.$$

Reorder the unselected covariates v_k such that for any unselected true covariate, $\tilde{\theta}_k > \tilde{\theta}_l$ whenever $l > k$. No additional orthogonalization is performed.

Apply a separate orthogonalization to $w_1, ..., w_m \in \mathsf{H}$. These are orthogonalized by the Gram-Schmidt process according to order of inclusion into $\widehat{S}$, with true covariates included (interspersed

according to when they were selected between the w_j) in the orthogonlization process. The resulting Gram-Schmidt-orthogonalized elements are renormalized to give

$$w_1, ..., w_m; v_1, ..., v_{s_0} \mapsto \tilde{w}_1, ..., \tilde{w}_m \in \mathsf{H}$$

such that the component of each $\tilde{w}_j$ orthogonal to the span of $\tilde{v}_1, ..., \tilde{v}_{s_0}$ in H has unit norm. This renormalization is possible whenever $\varphi_{\min}(m + s_0)(G_{\mathsf{H}}) > 0$. Therefore, $\tilde{w}_j$ can be decomposed into $\tilde{w}_j = \tilde{r}_j + \tilde{u}_j$ which satisfy $\tilde{r}_j \in \text{span}(\tilde{v}_1, ..., \tilde{v}_{s_0})$ and $\tilde{u}_j \in \text{span}(\tilde{v}_1, ..., \tilde{v}_{s_0})^{\perp}$ and $\|\tilde{u}_j\|_{\mathsf{H}} = 1$.

For ease of reading, the online supplemental material also presents additional descriptive notation and details about the orthogonalization constructions.

The two orthogonalizations are next related to each other by associating, for each $j \in A$, parameters

$$j \mapsto (\tilde{\gamma}_j \in \mathbb{R}^{s_0}, \tilde{\gamma}_{j\tilde{\varepsilon}} \in \mathbb{R})$$

which are defined as $\tilde{\gamma}_{jk} = \langle \tilde{w}_j, \tilde{v}_k \rangle_{\mathsf{H}}, \tilde{\gamma}_{j\tilde{\varepsilon}} = \langle \tilde{w}_j, \tilde{\varepsilon} \rangle_{\mathsf{H}}$. Assume without loss of generality that each component of $\tilde{\theta}$ is positive (the remainder of the proof does not depend on the sign assigned during orthogonalization). Similarly, assume without loss of generality that $\tilde{\gamma}_j' \tilde{\theta} \geqslant 0$.

A large part of the following analysis is at the level of the parameters $\tilde{\theta}, \tilde{\theta}_{\tilde{\varepsilon}}, \tilde{\gamma}_j, \tilde{\gamma}_{j\tilde{\varepsilon}}$. Therefore, some remarks are helpful. Aside from relating the two orthogonalizations, these parameters also encode information about incremental loss for various j and sets S. For example, in the case of Theorem 1, it can be shown that $-\Delta_j \ell(S) = \frac{1}{n} \frac{1}{\|\tilde{w}_j\|_{\mathsf{H}}^2} (\tilde{\theta}' \tilde{\gamma}_j + \tilde{\theta}_{\tilde{\varepsilon}} \tilde{\gamma}_{j\tilde{\varepsilon}})^2$ if S is the set of all covariates selected into $\widehat{S}$ before w_j. Similarly, $-\Delta_k \ell(S) = \frac{1}{n} \tilde{\theta}_k^2$ if S corresponds to the set $\{v_1, ..., v_{k-1}\}$. Relatedly, note that if $\tilde{\gamma}_j' \tilde{\theta}$ is large for sufficiently many j, then some dependence between $\tilde{\gamma}_j$ may be anticipated; see Tao (2014), for a general discussion of partial transitivity of correlation. This, however, heuristically creates tension with the fact that $\tilde{\gamma}_j$ arise from orthogonalized covariates.

A.3.2. Main Sparsity Bounds

Divide the set A of false covariates into two sets A_1 and A_2, with cardinalities m_1 and m_2, on the basis of the magnitude of $\tilde{\gamma}_{j\tilde{\varepsilon}}$. Set $A_1 = \left\{ j : |\tilde{\gamma}_{j\tilde{\varepsilon}}|^2 \leqslant \frac{t_{\mathsf{H}} n}{3\|\tilde{\varepsilon}\|_{\mathsf{H}}} \right\}$, $A_2 = A \setminus A_1$. Note that large values of $\tilde{\gamma}_{j\tilde{\varepsilon}}$ indicate higher dependence between orthogonalized versions of w_j and ε. Bounds on the size of A_1 are given first.

Let A_{1k} be the set of $j \in A_1$ such that j is selected prior to the k-th true selection, but not prior to any earlier true selections. For $j \in A_1, k, l \in S_0$, let $C_1 > 0$ and $1 \geqslant C_2 > 0$ be constants which satisfy

$$\tilde{\gamma}_j' \tilde{\theta} \geqslant \tilde{\theta}_k C_1 \text{ for } j \in A_{1k}, \text{ and } \tilde{\theta}_k \geqslant \tilde{\theta}_l C_2 \text{ for } l > k.$$

The two key constants C_1,C_2 encode information about *relative* incremental loss values at various points of the forward selection procedure. Lemma 5 calculates suitable C_1, C_2.

Lemma 5. *$\tilde{\gamma}_j' \tilde{\theta} \geqslant \tilde{\theta}_k \left(\frac{1}{6} \varphi_{\min}(m + s_0)(G_{\mathsf{H}}) \right)^{1/2}$ for $j \in A_{1k}$. In addition, in the case of Theorem 1, $\tilde{\theta}_k \geqslant \tilde{\theta}_l \varphi_{\min}(m + s_0)(G_{\mathsf{H}})^{1/2}$ for $l > k$. In the case of Theorem 4, $\tilde{\theta}_k \geqslant \tilde{\theta}_l \left(c''_{\text{test}} \varphi_{\min}(m + s_0)(G_{\mathsf{H}}) \right)^{1/2}$ for $l > k$.*

Define the following two $s_0 \times s_0$ matrices.

$$\Gamma : \ \Gamma_{kl} = \sum_{j \in A_{1k}} \tilde{\gamma}_{jl}, \quad B \ : \ \text{symmetric}, \ B_{kl} = \tilde{\theta}_l / \tilde{\theta}_k \text{ if } l \geqslant k.$$

Empty sums are taken to be 0. In case $\tilde{\theta}_k = 0$ for some k, $\tilde{\theta}_l / \tilde{\theta}_k$ is defined to be 1. The above definitions of Γ, B are useful because the diagonal elements of the product ΓB satisfy the equality $[\Gamma B]_{kk} = \sum_{j \in A_{1k}} \tilde{\gamma}_j' \tilde{\theta} / \tilde{\theta}_k$. This follows from Γ being upper triangular (by the orthogonalization construction) and from the fact that A_{1k} is empty if $\tilde{\theta}_k = 0$ (see remark in the proof of Lemma 5).

The definition of C_1 implies that $[\Gamma B]_{kk} \geqslant C_1 |A_{1k}|$ and subsequently

$$\mathrm{tr}(\Gamma B) \geqslant C_1 m_1.$$

The product ΓB has a convenient decomposition. Consider the set

$$\mathcal{G}_{s_0} = \left\{ Z \in \mathbb{R}^{s_0 \times s_0} : \begin{array}{c} Z_{kl} = \langle \mathsf{X}_k, \mathsf{Y}_l \rangle_{\mathsf{H}_1} \text{ for some elements } \ \|\mathsf{X}_k\|_{\mathsf{H}_1}, \|\mathsf{Y}_l\|_{\mathsf{H}_1} \leqslant 1 \\ \text{in some } s_0\text{-dimensional real Hilbert space } \mathsf{H}_1 \end{array} \right\}$$

Lemma 6. *The matrix product ΓB may be expressed as $\Gamma B = \Gamma C_3 \bar{Z}$ where C_3 is a constant which may be taken as $C_3 = C_2^{-2}$ and where $\bar{Z}' \in \mathcal{G}_{s_0}$.*

This decomposition is helpful because of the following result due to Grothendieck.

Lemma 7 (Grothendieck (1953))**.** $\sup_{Z \in \mathcal{G}_{s_0}} \mathrm{tr}(MZ) \leqslant K_G^{\mathbb{R}} \|M\|_{\infty \to 1}$.

Here, $K_G^{\mathbb{R}}$ is an absolute constant which is known to be less than 1.783. Importantly, it does not depend on s_0. The notation $\|\cdot\|_{\infty \to 1}$ indicates the operator norm for bounded linear operators $L^\infty \to L^1$. When the matrix M is $s_0 \times s_0$ dimensional, the implied L^∞, L^1 spaces are $L^\infty(\{1, ..., s_0\}), L^1(\{1, ..., s_0\})$ or equivalently $(\mathbb{R}^{s_0}, \|\cdot\|_\infty)$, $(\mathbb{R}^{s_0}, \|\cdot\|_1)$. The form used here is that described in Guédon and Vershynin (2016), Equation 3.2. Therefore,

$$C_3^{-1} C_1 m_1 \leqslant C_3^{-1} \mathrm{tr}(\Gamma B) = C_3^{-1} \mathrm{tr}(\Gamma C_3 \bar{Z}) = \mathrm{tr}(\Gamma' \bar{Z}') \leqslant K_G^{\mathbb{R}} \|\Gamma'\|_{\infty \to 1}.$$

In light of this lower bound on $\|\Gamma'\|_{\infty \to 1}$, there exists $\nu \in \{-1, 1\}^{s_0}$ such that

$$\|\Gamma' \nu\|_1 \geqslant \left(K_G^{\mathbb{R}}\right)^{-1} {C_3}^{-1} C_1 m_1.$$

On the other hand, $\|\Gamma' \nu\|_1$ may be upper bounded by a quantity that depends on s_0 by

$$\|\Gamma' \nu\|_1 \leqslant s_0^{1/2} \|\Gamma' \nu\|_2.$$

A key property of the $\tilde{\gamma}_j$, which constitute Γ, is that they are approximately orthogonal to each other in the sense of the following lemma. In particular, signed sums of $\tilde{\gamma}_j$ scale in norm like $m_1^{1/2}$ up to a factor depending on $\varphi_{\min}(m + s_0)(G_{\mathsf{H}})$.

Lemma 8. *For any signs $e_j \in \{-1, 1\}$, $\left\| \sum_{j \in A_1} e_j \tilde{\gamma}_j \right\|_2 \leqslant {m_1}^{1/2} \varphi_{\min}(m + s_0)(G_{\mathsf{H}})^{-1/2}$.*

Therefore, $\|\Gamma'\nu\|_2 = \|\sum_{k=1}^{s_0}\sum_{j\in A_{1k}}\nu_k\tilde{\gamma}_j\|_2 \leqslant m_1^{1/2}\varphi_{\min}(m+s_0)(G_{\mathsf{H}})^{-1/2}$, which, when combined with the bound $\|\Gamma'\nu\|_1 \leqslant s_0^{1/2}\|\Gamma'\nu\|_2$, immediately implies the following.

Lemma 9. $m_1 \leqslant \varphi_{\min}(m+s_0)(G_{\mathsf{H}})^{-1}C_1^{-2}{C_3}^2\left(K_G^{\mathbb{R}}\right)^2 s_0$.

Having controlled m_1, it is left to give a bound which controls m_2. The following lemma is proven by showing that the orthogonalization process $w_j \mapsto \tilde{w}_j$ cannot create too many variables, j, with large $\tilde{\gamma}_{j\tilde{\varepsilon}}$, given the relevant regularization condition is met.

Lemma 10. $m_2 \leqslant 3(m_1+s_0)$ *provided* $t^{1/2} \geqslant 2\varphi_{\min}(m+s_0)(G)^{-1}\|\mathbb{E}_n[x_i\varepsilon_i]\|_\infty$ *in the case of Theorem 1 and* $\mathrm{E}[\mathbb{E}_n[\varepsilon_i^{a2}]] \leqslant \frac{1}{2}\varphi_{\min}(m+s_0)(\mathrm{E}[G])^{-1}c'_{\text{test}}$ *in the case of Theorem 4.*

The final lemma restates the sparsity bounds of Theorems 1 and 4. Its proof involves only assembling the previous arguments. Recall that $m_1+m_2=m$ and that m is the number of false selections being considered.

Lemma 11. *In the case of Theorem 1, if* $t^{1/2} \geqslant 2\varphi_{\min}(m+s_0)(G)^{-1}\|\mathbb{E}_n[x_i\varepsilon_i]\|_\infty$ *holds, then also* $m \leqslant 80\times\varphi_{\min}(m+s_0)(G)^{-4}s_0$ *holds. In the case of Theorem 4,* $\widehat{s} \leqslant (80\times\varphi_{\min}(m+s_0)(\mathrm{E}[G])^{-4}c''^{-3}_{\text{test}}+1)s_0$.

This completes the proof of Theorems 1 and 4.

References

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

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

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

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.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014.

A. Belloni, V. Chernozhukov, C. Hansen, and D. Kozbur. Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4): 590–605, 2016.

D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *Annals of Statistics*, 44(2):813–852, 04 2016.

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

P. Bühlmann. Boosting for high-dimensional linear models. *Annals of Statistics*, 34(2):559–583, 2006.

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

E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated Lasso. *ArXiv e-prints*, 2016.

A. Chudik, G. Kapetanios, and M. H. Pesaran. A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica*, 86 (4):1479–1512, 2018.

A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. ACM, 2011.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, 32(2):407–451, 2004.

M. Efroymson. Stepwise regression–a backward and forward look. *Florham Park, New Jersey*, 1966.

C. Elbers, J. O. Lanjouw, and P. Lanjouw. Micro–level estimation of poverty and inequality. *Econometrica*, 71(1):355–364, 2003.

A. Fiszbein, N. Schady, F. Ferreira, M. Grosh, N. Keleher, P. Olinto, and E. Skoufias. *Conditional Cash Transfers: Reducing Present and Future Poverty*. The World Bank, 2009.

W. Fithian, J. Taylor, R. Tibshirani, and R. Tibshirani. Selective Sequential Model Selection. *ArXiv e-prints*, Dec. 2015.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In L. Saitta, editor, *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)*, pages 148–156. Morgan Kaufmann, 1996. ISBN 1-55860-419-7.

A. Grothendieck. Resume de la theorie metrique des produits tensoriels topologiques. *Boletim Sociedade de Matemática de São Paulo*, 8:1 – 79, 1953.

M. G. G'Sell, S. Wager, A. Chouldechova, and R. Tibshirani. Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2):423–444, 2016.

O. Guédon and R. Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165, 2016.

R. Hanna and B. A. Olken. Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives*, 32(4):201–26, 2018.

T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2009.

T. Hastie, R. Tibshirani, and T. R. Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso. *ArXiv e-prints*, July 2017.

J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Ann.*

Statist., 38(4):2282–2313, 2010. ISSN 0090-5364.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

B.-Y. Jing, Q.-M. Shao, and Q. Wang. Self-normalized Cramèr-type large deviations for independent random variables. *Annals of Probability*, 31(4):2167–2215, 2003.

D. Kozbur. Sharp Convergence Rates for Forward Regression in High-Dimensional Sparse Linear Models. *ArXiv e-prints*, Feb. 2017a.

D. Kozbur. Testing-based forward model selection. *American Economic Review*, 107(5):266–69, May 2017b.

H. Leeb and B. M. Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376, 2008. ISSN 0266-4666. .

A. Li and R. F. Barber. Accumulation tests for fdr control in ordered hypothesis testing. *Journal of the American Statistical Association*, 112(518):837–849, 2017.

Y. Luo and M. Spindler. L2 Boosting for Economic Applications. *ArXiv e-prints*, Feb. 2017.

N. Meinshausen and P. Buhlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

W. K. Newey and K. D. West. A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

A. Nichols and L. McBride. Retooling poverty targeting using out-of-sample validation and machine learning. *World Bank Economic Review*, 32(3):531–550, 2018.

M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:10251045, 2008.

M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59:3434–3447, 2013.

T. Tao. When is correlation transitive? url: https://terrytao.wordpress.com/2014/06/05/when-is-correlation-transitive/, June 2014.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact Post-Selection Inference for Sequential Regression Procedures. *ArXiv e-prints*, Jan. 2014.

J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, Oct 2004. ISSN 0018-9448.

S. van de Geer, P. Bhlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

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

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

C.-H. Zhang and S. 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.

T. Zhang. On the consistency of feature selection using greedy least squares. *Journal of Machine Learning*, 10:555–568, 2009.

T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57(7):4689–4708, 2011.