Testing Many Zero Restrictions in a High Dimensional Linear Regression Setting

Jonathan B. Hill*

Dept. of Economics, University of North Carolina, U.S.

This draft: December 11, 2023

Abstract

We propose a test of many zero parameter restrictions in a high dimensional linear iid regression model with $k >> n$ regressors. The test statistic is formed by estimating key parameters one at a time based on many low dimension regression models with nuisance terms. The parsimoniously parametrized models identify whether the original parameter of interest is or is not zero. Estimating fixed low dimension sub-parameters ensures greater estimator accuracy, it does not require a sparsity assumption nor therefore a regularized estimator, it is computationally fast compared to, e.g., de-biased Lasso, and using only the largest in a sequence of weighted estimators reduces test statistic complexity and therefore estimation error. We provide a parametric wild bootstrap for p-value computation, and prove the test is consistent and has non-trivial $\sqrt{n/\{\ln(n)\mathcal{M}_n\}}$-local-to-null power where $\mathcal{M}_n$ is the l_∞ covariate fourth moment.

Key words and phrases: high dimensional inference, max-test, linear regression

AMS classifications : 62G10, 62M99, 62F35.

JEL classifications : C12, C55.

1 Introduction

Regression settings where the number k of covariates x_t may be much larger than the sample size n ($k >> n$) is a natural possibility in cross sections, panels and spatial settings where an enormous amount of information is available. In economics and finance, for example,

*Department of Economics, University of North Carolina, Chapel Hill, North Carolina, U.S. E-mail:jbhill@email.unc.edu; https://jbhill.web.unc.edu. This paper was previously circulated under the title "*Testing (Infinitely) Many Zero Restrictions*".

reasons include the size of surveys (e.g. U.S. Census, Current Population Survey, National Longitudinal Survey of Youth), increasingly sophisticated survey techniques (e.g. text records and word counts, household transaction data), the unit of observation (e.g. county wide house prices are dependent on neighbor county prices), use of dummy variables for groups (e.g. age, race, region, etc.), and panel data fixed effects. See, for instance, Fan, Lv, and Qi (2011) and Belloni, Chernozhukov, and Hansen (2014) for recent surveys. In non-social sciences high dimensionality arises with, for instance, genomic data and network and communication data.[1]

Inference in this case is typically done post estimation based on a regression or moment condition model with an imposed sparsity condition, or related structural assumption. Sparsity in practice is achieved by a shrinkage or regularized estimator, while recent theory allows for an increasing number of candidate and relevant covariates as the sample size increases (see, e.g., Fan and Peng, 2004; Huang, Horowitz, and Ma, 2008). Valid inference, however, typically exists only for the non-zero valued parameter subset, although there is a nascent literature for gaining inference on the (number of) zero valued parameters (e.g. Carpentier and Verzelen, 2019). Conversely, Bühlmann (2013) develops a test of zero restrictions for a high dimensional linear model with fixed design $\{x_1, ..., x_n\}'$ using bias-corrected ridge regression. Their p-value construction, however, is based on a linear model with Gaussian errors and an upper bound on the distribution of the ridge estimator. Low dimension projection methods offer another solution (e.g. Zhang and Zhang, 2014). Imposing sparsity at all, however, may be too restrictive, in particular without justification via a pre-estimation test procedure.

Dezeure, Bühlmann, and Zhang (2017) work in the de-biased Lasso [DBL] setting of Zhang and Zhang (2014) to deliver a useful post-estimation max-t-test under a shrinkage assumption. The max-test in Dezeure, Bühlmann, and Zhang (2017) is similar to ours in the sense that they bootstrap the maximum t-statistic over an arbitrary k-set of t-statistics, as long as $\ln(k) = o(n^{1/7})$ under covariate boundedness. Boundedness, however, is unlikely to exist in key set-

[1]As a testament to the increasing amount of available data, in November 2022 the 27th General Conference on Weights and Measures presented new numerical prefixes handling, among other magnitudes, 10^{30}.

tings in economics, including nonlinear models involving propensity scores, inverse probability weighting, and log-linearization; and many standard models in finance like ARMA-GARCH. Although we work in an iid linear setting to focus ideas, our methods can only reasonably, and usefully, be extended to time series models under unboundedness.

Consider a sample of iid observations $\{x_t, y_t\}_{t=1}^n$ with scalar y_t, and covariates or design points $x_t = [x'_{\delta,t}, x'_{\theta,t}]'$, and a linear regression model for simplicity and to fix ideas:

$$y_t = \delta'_0 x_{\delta,t} + \theta'_0 x_{\theta,t} + \epsilon_t = \beta'_0 x_t + \epsilon_t. \tag{1}$$

This paper presents a test of $H_0 : \theta_0 = 0$ vs. $H_1 : \theta_{0,i} \neq 0$ for at least one i. The nuisance parameter δ_0 has a fixed dimension $k_\delta < n$, but θ_0 may have dimension $k_\theta >> n$. We assume $k_\theta \to \infty$, but $k_\theta \to k_\theta^* \in \mathbb{N}$ can be easily allowed. In order to show that k_θ depends on n we write $k_{\theta,n} = k_\theta$ throughout. Let β_0 be the unique minimizer of the squared error loss $E[(y_t - \beta' x_t)^2]$, where $E[\epsilon_t] = 0$. We do not require (ϵ_t, x_t) to be mutually independent, but use the second order condition $P(E[\epsilon_t^2 | x_t] = \sigma^2) = 1$ for some finite $\sigma^2 > 0$ to focus ideas. Thus (δ_0, θ_0) should be seen as being possibly pseudo-true in the sense of Sawa (1978) and White (1982). Heterogeneity, including (conditional) heteroscedasticity, is straightforward to include, but omitted here due to space considerations. See the appendix for all assumptions. We can test general hypotheses $H_0 : \theta_0 = \tilde{\theta}$ vs. $H_1 : \theta_0 \neq \tilde{\theta}$, but testing $\theta_0 = 0$ naturally saves notation.

Although $k_{\theta,n} >> n$, we do not require a sparsity condition common in the high dimensional estimation literature (see, e.g., Hastie, Tibshirani, and Friedman, 2009; Zhang and Zhang, 2014; Dezeure, Bühlmann, and Zhang, 2017). There are many uses for (1), including panels with many fixed effects, estimation with many instrumental variables, and linear projections of nonlinear response, e.g. series expansions from flexible functional forms in machine learning (e.g. Andrews, 1991; Royston and Altman, 1994; Belloni, Chernozhukov, Fernandez-Val, and Hansen, 2017; Cattaneo, Jansson, and Newey, 2018; Gautier and Rose, 2021).

There is a large literature on high dimensional linear regression with an increasing number of covariates; consult Zhang and Zhang (2014), Dezeure, Bühlmann, and Zhang (2017), Cattaneo, Jansson, and Newey (2018) and Li and Müller (2021), amongst others. This literature concerns estimation for high dimensional models, with inference *post*-estimation. We, however, approach inference on θ_0 *pre*-estimation, in the sense that (1) is not estimated, hence a penalized or regularized estimator like (de-biased) Lasso, a projection method, or partialing out parameters are not required. Instead, we operate on many low dimension regression models, each with one key parameter $\theta_{0,i}$ to be tested, and any other nuisance parameters δ_0 to be estimated. Parsimony improves estimation of each $\theta_{0,i}$ when $k_{\theta,n}$ is large, and allows us to sidestep sparsity considerations under H_1. Indeed, pre-estimation itself allows one to test a sparsity assumption based on a target set of parameters.

Our test approach is as follows. Write $\beta_{(i)} \equiv [\delta', \theta_i]'$, and construct parsimonious models:

$$y_t = \delta_{(i)}^{*\prime} x_{\delta,t} + \theta_i^* x_{\theta,i,t} + v_{(i),t} = \beta_{(i)}^{*\prime} x_{(i),t} + v_{(i),t}, \;\; i = 1, ..., k_{\theta,n}, \tag{2}$$

where (pseudo-true) $\beta_{(i)}^*$ for each i satisfy $E[(y_t - \beta_{(i)}^{*\prime} x_{(i),t}) x_{(i),t}] = 0$. Thus the i^{th} model corresponds to a single unique regressor $x_{\theta,i,t}$ from $x_{\theta,t} = [x_{\theta,i,t}]_{i=1}^{k_{\theta,n}}$, and nuisance regressor subset $x_{\delta,t}$. The methods proposed here can be directly extended to more general settings, including nonlinear regression models with non-identically distributed observations, conditional moment models, random choice and quantile regression to name a few.

A key result shows that under mild conditions, $\theta^* \equiv [\theta_i^*]_{i=1}^{k_{\theta,n}} = 0$ *if and only if* $\theta_0 = 0$, thus each nuisance parameter $\delta_{(i)}^*$ is identically the true δ_0. Obviously we cannot generally identify θ_0 (or δ_0) under the alternative, but the set of models in (2) can identify whether $\theta_0 = 0$ is true or not. Thus (2) represents the set of *least parameterized* models, and therefore the fewest models ($k_{\theta,n}$), required to identify whether the null hypothesis is correct, leading to an efficiency gain.

Now let $\hat{\beta}_{(i)} \equiv [\hat{\delta}'_{(i)}, \hat{\theta}_i]'$ minimize the parsimonious least squares loss $1/n \sum_{t=1}^{n}(y_t - \beta'_{(i)}x_{(i),t})^2$. The test statistic is the normalized maximum weighted $\hat{\theta}_i$:

$$\mathcal{T}_n = \max_{1 \leq i \leq k_{\theta,n}} \left| \sqrt{n}\mathcal{W}_{n,i}\hat{\theta}_i \right|,$$

where $\{\mathcal{W}_{n,i}\}_{n \geq 1}$ are sequences of possibly stochastic weights, $\mathcal{W}_{n,i} > 0$ *a.s.* for each i, with non-random (probability) limits $\mathcal{W}_i \in (0, \infty)$. The max-test rejects H_0 at level α when an appropriate p-value approximation $\hat{p}_n$ satisfies $\hat{p}_n < \alpha$. The weights $\mathcal{W}_{n,i}$ allow for a variety of test statistics, including $\max_{1 \leq i \leq k_{\theta,n}} |\sqrt{n}\hat{\theta}_i|$, or a max-t-statistic. Bootstrap inference for the max-statistic does not require a covariance matrix inversion like Wald and LM tests, and operates like a shrinkage estimator by using only the most relevant (weighted) estimator. Our high dimensional parametric wild bootstrap theory exploits multiplier bootstrap theory in Chernozhukov, Chetverikov, and Kato (2013), and the concept of *weak convergence in probability* in Gine and Zinn (1990). The max-test is asymptotically correctly sized, and consistent because $k_{\theta,n} \to \infty$. Bootstrapping in high dimension has a rich history, e.g. Bickel and Freedman (1983), Mammen (1993), Chernozhukov, Chetverikov, and Kato (2013, 2017) and Dezeure, Bühlmann, and Zhang (2017) and their references.

We detail four covariate cases in order to bound $k_{\theta,n}$, yielding $\ln(k_{\theta,n}) = o(n^{1/7})$ under (i) boundedness, (ii) sub-Gaussianicity, or (iii) sub-exponentiality; and (iv) under $\mathcal{L}_p$-boundedness $o((n/\ln(n)^2)^{p/8})$ where $p \geq 4$. Furthermore, the max-test has non-trivial power against a sequence of $\sqrt{n/\{\ln(k_{\theta,n}) E[\max_{1 \leq i \leq k_{\theta,n}} |x_{(i),t}|^4]\}}$-local alternatives.

The dimension k_δ of δ_0 is assumed fixed and $k_\delta < n$ to focus ideas, but can in principle vary with n, or be unbounded. In a more general framework, if δ_0 is sparse then a penalized estimator can be used to estimate $[\delta^*_{(i)}, \theta^*_i]$, like DBL, ridge or Dantzig. If $k_\delta \to \infty$ as $n \to \infty$ then the nuisance variables can be partialed out, similar to Cattaneo, Jansson, and Newey (2018). If δ_0 is a (infinite dimensional) function, e.g. $y_t = \delta_0(x_{\delta,t}) + \theta'_0 x_{\theta,t} + \epsilon_t$, then $[\delta^*_{(i)}, \theta^*_i]$

can be estimated by sieves (e.g. Chen and Pouzo, 2015). In our setting inference is on the high dimensional θ_0, contrary to Chernozhukov et al (2016) and Cattaneo, Jansson, and Newey (2018) who test a fixed low dimension parameter after partialing out a high dimensional term.

Conventional test statistics typically exhibit size distortions due both to poor estimator sample properties, and from inverting a high dimension covariance matrix estimator that may be a poor approximation of the small or asymptotic variance. The challenge of high dimensional covariance matrix estimation is well documented (e.g. Chen, Xu, and Wu, 2013). A bootstrap method is therefore typically applied, but bootstrap tests may only have size corrected power equal to the original test, which may be low under parameter proliferation (e.g. Davidson and MacKinnon, 2006; Ghysels, Hill, and Motegi, 2020). Indeed, the bootstrapped Wald test when $k_{\theta,n} + k_{\delta} < n$ is undersized with low power in linear iid models when k_{δ} is even mildly large (e.g. $k_{\delta} = 10$)[2], and potentially significantly under-sized with low power when $k_{\theta,n}$ is large, becoming acute when there are nuisance parameters. Max-tests yield better (typically sharp) size, while power improvements over the Wald test can be sizable when deviations from the null are small.

Reduced dimension regression models in the high dimensional parametric statistics and machine learning literatures are variously called *marginal regression*, *correlation learning*, and *sure screening* (e.g. Fan and Lv, 2008; Genovese, Jin, Wasserman, and Yao, 3023), and includes *canonical correlation analysis* [CCA] (e.g. McKeague and Zhang, 2022). McKeague and Qian (2015), for example, regress y against one covariate x_i at a time, $i = 1, ..., k < \infty$, and compute the most relevant index $\hat{k} \equiv \arg\max_{1 \leq i \leq k} |\hat{\theta}_i|$ where $\hat{\theta}_i \equiv \widehat{cov}(y, x_i)/\widehat{var}(x_i)$, and test $\tilde{H}_0 : \theta_0 = [cov(y, x_i)/var(x_i)]_{i=1}^{k} = 0$ using $\hat{\theta}_{\hat{k}}$. Thus they work outside a general regression setting, consider only a fixed number k of possible covariates, do not consider a subset of regressor predictability (e.g. $x_{\theta,t}$) with other nuisance parameters (δ_0), and use linear dependence for defining predictability. Related projections to low dimension settings are considered

[2]This may carry over to (semi)nonparametric settings with fixed low dimension δ and infinite dimensional function θ_0. See Chen and Pouzo (2015) for simulation evidence of sieve based bootstrapped t-tests.

in Bühlmann (2013), Zhang and Zhang (2014), and Ghysels, Hill, and Motegi (2020), amongst others. McKeague and Zhang (2022) operate on the canonical correlation between bounded high dimensional variables $x_t \in \mathbb{R}^p$ and $y_t \in \mathbb{R}^q$. They propose nonparametric max-type tests of the global null hypothesis that there are no linear relationships between any subsets of x_t and y_t. In economics and finance settings, however, there will typically be nuisance covariates to control for, and the potential for nonlinear relationships.

Compared to McKeague and Qian (2015) and McKeague and Zhang (2022), we also work pre-estimation, but within a high dimensional regression model that allows for misspecification and nuisance parameters, and test the parameter subset $H_0 : \theta_0 = 0$ where θ_0 may have a diverging dimension. One could train y_t on just the key subset $x_{\theta,t}$, cf. McKeague and Zhang (2022), but that neglects intermediary relationships (or controls) with $x_{\delta,t}$, and they impose boundedness on both x_t and y_t. We also do not impose sparsity (under H_1), nor any distributional assumptions, other than bounding $k_{\theta,n}$ under four covariate cases from boundedness to $\mathcal{L}_p$-boundedness with $p \geq 4$. If there are no nuisance parameters ($k_\delta = 0$) then our theory should allow for a high dimensional expansion of McKeague and Qian's (2015) method, or an extension allowing for partialled out nuisance parameter, but we leave that arc for future consideration.

Moreover, our method has a straightforward extension to nonlinear models and a broad class of extremum estimators so that predictability may not align with linear dependence between y_t and the $x_{i,t}^{\prime}s$ (as opposed to correlations, cf. McKeague and Qian (2015) and McKeague and Zhang (2022)). Consider a nonlinear response $f(x_t, \beta_0)$ with known f and parsimonious versions $f(x_t, [\delta_{(i)}^*; 0, ..., \theta_{(i)}^*, ...0])$, where $[0, ..., \theta_{(i)}, ...0]'$ is a $k_{\theta,n} \times 1$ zero vector with $\theta_{(i)}$ in the i^{th} row. Thus θ_0 need not align one-for-one with a regressor subset $x_{\cdot,t}$. Examples abound in the financial and macro-econometrics, neural network and machine learning literatures, including switching models, flexible functional forms, and basis expansions.

Forward stepwise regressions exploit parsimonious models with increasing dimension. In the estimation literature low dimension, typically linear, models are built by adding one covariate at a time (e.g. Amemiya, 1980; Stone, 1981; Royston and Altman, 1994; Castle, Qin, and Reed, 2009). Stepwise regression, like regularized estimators, depend on a reliable choice of tuning parameter (or stopping rule). Our approach is different because we do not build up a model; rather, we use a different class of parsimonious models that combines just one key covariate with the nuisance covariates. Our models are true under the null, whereas stepwise methods typically do not hinge on a presumed hypothesis. Classically, Kabe (1963) tests whether a low dimension subset $\theta_0 = 0$ with stepwise regressions that partial out δ_0, effectively using parsimonious models. This idea has been treated many times since, e.g. Cattaneo, Jansson, and Newey (2018) and Chernozhukov et al (2016). Our method does not partial out low or high dimensional components, nor add covariates one at time based on predictor power.

There are several differences between our method and the max-test in Dezeure, Bühlmann, and Zhang (2017). First, they consider different bootstrap techniques designed to allow for heteroscedastic errors. As discussed above, we focus on homoscedastic errors given space constraints. Second, we do not estimate (1) directly, hence we do not require shrinkage under either hypothesis. Third, we allow for any covariate as long as it has a fourth moment, and derive bounds on $k_{\theta,n}$ under four covariate cases. Fourth, we provide a class of weighted max-statistics covering $\max_{1 \leq i \leq k_{\theta,n}} |\sqrt{n}\hat{\theta}_i|$ as well as a max-t-test. Fifth, computation time for our bootstrapped p-value is *significantly* faster than required for the bootstrapped node-wise DBL because we do not require a regularized estimator.[3] For further reading on post-estimation hypothesis testing for regularized estimators under sparsity, see Liu and Luo (2014), and see Cai and Guo (2017) and Cai, Guo, and Xia (2023) for minimax optimality properties and

[3]Using Matlab (with coordinate descent and ADMM algorithms) via a SLURM scheduler on the Longleaf cluster at UNC, with 128 workers on 1 node, and setting $n = 100$, $k_\delta = 0$, $k_{\theta,n} = 200$, and 1000 bootstrap samples, *one* bootstrapped p-value under H_0 for the DBL max-statistic with 5-fold cross-validation took 335 seconds (5.6 minutes). Increase $k_{\theta,n}$ to 480, as in our simulation study, yielded a computation time of 8109 seconds (135.2 minutes). The bootstrapped parsimonious max-test, by comparison, took only 4.17 and 5.1 seconds, respectively, yielding up to a $1600x$ computation time gain. Increase n to 250 and set $k_{\theta,n} = 1144$, and DBL required 50.21 hours, while our method took just 5.2 seconds (roughly $35,000x$ faster). See Section 4 for complete simulation details.

recent surveys.

The remainder of this paper proceeds as follows. Section 2 covers hypothesis identification and test statistic asymptotics, and Section 3 presents bootstrap p-value theory. A Monte Carlo study is presented in Section 4, and parting comments are left for Section 5. Assumptions, proofs of main results, and a subset of simulation results are relegated to the appendix. The supplemental material [SM] Hill (2023) contains omitted proofs and all simulation results.

We assume all random variables exist on a complete measure space. $|x| = \sum_{i,j} |x_{i,j}|$ is the l_1-norm, $|x|_2 = (\sum_{i,j} x_{i,j}^2)^{1/2}$ is the Euclidean or l_2 norm, and $||A|| = \max_{|\lambda|_2}\{|A\lambda|_2/|\lambda|_2\}$ is the spectral norm for finite dimensional square matrices A (and the Euclidean norm for vectors). $||\cdot||_p$ denotes the L_p-norm. *a.s.* is *almost surely*. $\mathbf{0}_k$ denotes a zero vector with dimension $k \geq 1$. Write r-vectors as $x \equiv [x_i]_{i=1}^r$. $[\cdot]$ rounds to the nearest integer. $K > 0$ is non-random and finite, and may take different values in different places. $awp1$ = asymptotically with probability approaching one. We say z has *sub-exponential* distribution tails when $P(|z| > \varepsilon) \leq b\exp\{c\varepsilon\}$ for some $(b,c) > 0$ and all $\varepsilon > 0$ (see, e.g., Vershynin, 2018, Chap.2.7).

2 Max-Test and p-Value Computation

We first show that use of the set of parsimonious models $y_t = \delta_{(i)}^{*\prime} x_{\delta,t} + \theta_i^* x_{\theta,i,t} + v_{(i),t}$, $i = 1,...,k_{\theta,n}$, may be used to construct a test of $\theta_0 = \mathbf{0}_{k_{\theta,n}}$. Collect all θ_i^* in (2) into $\theta^* \equiv [\theta_i^*]_{i=1}^{k_{\theta,n}}$.

Theorem 2.1. $\theta_0 = \mathbf{0}_{k_{\theta,n}}$ *if and only if* $\theta^* = \mathbf{0}_{k_{\theta,n}}$. *Under* $H_0 : \theta_0 = \mathbf{0}_{k_{\theta,n}}$ *hence* $\delta_{(i)}^* = \delta_0$ $\forall i$, *and under* $H_1 : H_1 : \theta_{0,i} \neq 0$ *for at least one* $i \in \{1,...,k_{\theta,n}\}$ *there exists an* i *such that* $\theta_i^* \neq 0$.

2.1 Max-Test Gaussian Approximation

Now define parsimonious least squares loss gradient and Hessians,

$$\widehat{\mathcal{G}}_{(i)} \equiv -\frac{1}{n}\sum_{t=1}^{n} v_{(i),t}x_{(i),t},\ \widehat{\mathcal{H}}_{(i)} \equiv \frac{1}{n}\sum_{t=1}^{n} x_{(i),t}x_{(i),t}^{\prime},\ \ \mathcal{H}_{(i)} \equiv E\left[x_{(i),t}x_{(i),t}^{\prime}\right],$$

and define the usual least squares first order terms: $\hat{\mathcal{Z}}_{(i)} \equiv -\sqrt{n}\mathcal{H}_{(i)}^{-1}\widehat{\mathcal{G}}_{(i)} = \mathcal{H}_{(i)}^{-1}n^{-1/2}\sum_{t=1}^{n} v_{(i),t}x_{(i),t}$. Under H_0 of course $v_{(i),t} = \epsilon_t$ for each i by Theorem 2.1. By standard least squares arguments $0 = \widehat{\mathcal{G}}_{(i)} + \widehat{\mathcal{H}}_{(i)}(\hat{\beta}_{(i)} - \beta_{(i)}^{*})$, hence

$$\sqrt{n}\left(\hat{\beta}_{(i)} - \beta_{(i)}^{*}\right) = -\sqrt{n}\mathcal{H}_{(i)}^{-1}\widehat{\mathcal{G}}_{(i)} - \left\{\widehat{\mathcal{H}}_{(i)}^{-1} - \mathcal{H}_{(i)}^{-1}\right\}\sqrt{n}\widehat{\mathcal{G}}_{(i)} \equiv \hat{\mathcal{Z}}_{(i)} + \hat{\mathcal{R}}_{i}.$$

We want a high dimensional asymptotic Gaussian approximation $|\max_{1\leq i\leq k_{\theta,n}}|\sqrt{n}\mathcal{W}_{n,i}\hat{\theta}_i| - \max_{1\leq i\leq k_{\theta,n}}|\mathcal{W}_i\boldsymbol{Z}_{(i)}|| \overset{p}{\rightarrow} 0$ where $\{\boldsymbol{Z}_{(i)}\}_{i\in\mathbb{N}}$ is Gaussian. Set all weights $\mathcal{W}_{n,i} = 1$ to ease notation for now. Then:

$$\begin{aligned}&\left|\max_{1\leq i\leq k_{\theta,n}}\left|\sqrt{n}\hat{\theta}_i\right| - \max_{1\leq i\leq k_{\theta,n}}\left|\boldsymbol{Z}_{(i)}\right|\right| \\ &\qquad\leq \left|\max_{1\leq i\leq k_{\theta,n}}\left|[\mathbf{0}_{k_\delta}', 1]\hat{\mathcal{Z}}_{(i)}\right| - \max_{1\leq i\leq k_{\theta,n}}\left|\boldsymbol{Z}_{(i)}\right|\right| + \max_{1\leq i\leq k_{\theta,n}}\left|[\mathbf{0}_{k_\delta}', 1]\hat{\mathcal{R}}_{i}\right|.\end{aligned} \tag{3}$$

We use a now standard Gaussian approximation theory under H_0 to bound $\max_{1\leq i\leq k_{\theta,n}}|[\mathbf{0}_{k_\delta}', 1]\hat{\mathcal{Z}}_{(i)}| - \max_{1\leq i\leq k_{\theta,n}}|\boldsymbol{Z}_{(i)}|$, and high dimensional concentration (like) inequalities to bound $\max_{1\leq i\leq k_{\theta,n}}|[\mathbf{0}_{k_\delta}', 1]\hat{\mathcal{R}}_{i}|$. If, however, we have stochastic weights $\{\mathcal{W}_{n,i}\}$ then we must contend with $\max_{1\leq i\leq k_{\theta,n}}|\mathcal{W}_{n,i} - \mathcal{W}_i|$ by assumption or by case, as we below.

Write $\mathcal{M}_n \equiv E[\max_{1\leq i\leq k_{\theta,n}}|x_{(i),t}|^4]$. We have the following general result proving $\max_{1\leq i\leq k_{\theta,n}}|\hat{\mathcal{R}}_i| \overset{p}{\rightarrow} 0$ provided $\ln(k_{\theta,n}) = o(\sqrt{n}/\mathcal{M}_n)$. Here and in the sequel we explore four broad covariate cases in order to bound $\mathcal{M}_n$ (and thereby bound $k_{\theta,n}$), in each case uniformly in i:

$$\begin{array}{llll} \boldsymbol{(i)} & \text{bounded } x_{(i),t} & \boldsymbol{(ii)} & \text{sub-Gaussian } \left|x_{(i),t}\right|^4 \\ \boldsymbol{(iii)} & \text{sub-exponential } \left|x_{(i),t}\right|^4 & \boldsymbol{(iv)} & \mathcal{L}_p\text{-bounded } x_{(i),t},\ p \geq 4 \end{array} \tag{4}$$

Examples of (ii) and (iii) are, respectively, $x_{(i),t} = \ln|y_{(i),t}|$ or $x_{(i),t} = |y_{(i),t}|^{1/4}$ for sub-Gaussian $y_{(i),t}$; and $x_{(i),t} = |y_{(i),t}|^{1/2}$ for sub-Gaussian $y_{(i),t}$.

Following standard high dimensional concentration theory, cf. Nemirovski's inequality, the thinner the tails then the greater the allowed upper bound on $k_{\theta,n}$. Assumption 1 is discussed in Appendix A.1. Proofs of lemmas are presented in SM.

Lemma 2.2. *Let H_0 and Assumption 1 hold. Let the weight sequences $\{\mathcal{W}_{n,i}\}_{n\in\mathbb{N}}$ satisfy $\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_{n,i} - \mathcal{W}_i| = O_p(\sqrt{\ln(k_{\theta,n})\mathcal{M}_n/n})$ for non-stochastic $\mathcal{W}_i \in (0,\infty)$. We have:*

$$\max_{1\leq i\leq k_{\theta,n}} \left|[\mathbf{0}'_{k_\delta}, 1]\hat{\mathcal{R}}_i\right| = O_p\left(\ln(k_{\theta,n})\mathcal{M}_n/\sqrt{n}\right). \tag{5}$$

Then $\max_{1\leq i\leq k_{\theta,n}} |[\mathbf{0}'_{k_\delta}, 1]\hat{\mathcal{R}}_i| \xrightarrow{p} 0$ if under covariate case (i) $\ln(k_{\theta,n}) = o(\sqrt{n})$; (ii) $\ln(k_{\theta,n}) = o(n^{1/3})$; (iii) $\ln(k_{\theta,n}) = o(n^{1/4})$; or (iv) $k_{\theta,n} = o((n/\ln(n)^2)^{p/8})$ where $p \geq 4$.

Remark 2.1. The weight limit $\mathcal{W}_i$ is assumed non-random because the Gaussian approximation theory below requires $\mathcal{W}_i\boldsymbol{Z}_{(i)}(\cdot)$ to be Gaussian.

Remark 2.2. The presence of the l_∞ fourth moment $\mathcal{M}_n$ arises from dealing with $\max_{1\leq i\leq k_{\theta,n}} ||\widehat{\mathcal{H}}_{(i)}^{-1} - \mathcal{H}_{(i)}^{-1}||$ in a high dimensional least squares environment, cf. Nemirovski's moment inequality (e.g. Bühlmann and Van De Geer, 2011, Lemma 14.24).

Remark 2.3. The $k_{\theta,n}$ bounds are more lenient than for the Gaussian approximation, below, except under $\mathcal{L}_p$-boundedness. Ultimately this is due to the relative lack of complexity that is required to prove $\max_{1\leq i\leq k_{\theta,n}} |[\mathbf{0}'_{k_\delta}, 1]\hat{\mathcal{R}}_i| \xrightarrow{p} 0$ via Nemirovski's inequality, compared to a Gaussian approximation theory for $\max_{1\leq i\leq k_{\theta,n}} |[\mathbf{0}'_{k_\delta}, 1]\hat{\mathcal{Z}}_{(i)}|$ (Chernozhukov, Chetverikov, and Kato, 2013, cf.). Thus, in practice the imperative bounds on $k_{\theta,n}$ under cases (i)-(iii) appear in Theorem 2.4 below.

The weight approximation $\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_{n,i} - \mathcal{W}_i| = O_p(\sqrt{\ln(k_{\theta,n})\mathcal{M}_n/n})$ seems unavoidable, but is justified by noting that it holds (trivially) with flat $\mathcal{W}_{n,i} = 1$, or inverted standard errors. The latter is formally stated as follows. Define residuals $\hat{v}_{(i),t} = y_t - \hat{\beta}'_{(i)}x_{(i),t}$. Squared standard errors for $\sqrt{n}\hat{\theta}_i$ are $\widehat{\mathcal{S}}_{(i)}^2 \equiv [\widehat{\mathcal{H}}_{(i)}^{-1}]_{1,1}\hat{\mathcal{V}}_{(i),n}^2$ where $\hat{\mathcal{V}}_{(i),n}^2 = 1/n\sum_{t=1}^n \hat{v}_{(i),t}^2$, and the asymptotic value is $\mathcal{S}_{(i)}^2 \equiv [\mathcal{H}_{(i)}^{-1}]_{1,1}E[v_{(i),t}^2]$. Now write $\mathcal{W}_{n,i} = 1/\widehat{\mathcal{S}}_{(i)}$ and $\mathcal{W}_i = 1/\mathcal{S}_{(i)}$. The following result with the usual Slutsky theorem argument yields in general $\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_{n,i} - \mathcal{W}_i| = O_p(\sqrt{\ln(k_{\theta,n})\mathcal{M}_n/n})$.

Lemma 2.3. *Let Assumption 1 hold. Then* $\max_{1\leq i\leq k_{\theta,n}} |\widehat{\mathcal{S}}^2_{(i)} - \mathcal{S}^2_{(i)}| = O_p(\sqrt{\ln(k_{\theta,n})\mathcal{M}_n/n})$.

Now with $\hat{\mathcal{Z}}_{(i)} = -\sqrt{n}\mathcal{H}^{-1}_{(i)}\widehat{\mathcal{G}}_{(i)}$ we derive by case sequences $\{k_{\theta,n}\}$, and scalar normal random variables $\boldsymbol{Z}_{(i)}(\lambda) \sim N(0, E[\epsilon_t^2]\lambda'\mathcal{H}^{-1}_{(i)}\lambda)$, such that for each $\lambda \in \mathbb{R}^{k_\delta+1}$, $\lambda'\lambda = 1$, the Kolmogorov distance

$$\rho_n(\lambda) \equiv \sup_{z\geq 0}\left|P\left(\max_{1\leq i\leq k_{\theta,n}}\left|\lambda'\hat{\mathcal{Z}}_{(i)}\right| \leq z\right) - P\left(\max_{1\leq i\leq k_{\theta,n}}\left|\boldsymbol{Z}_{(i)}(\lambda)\right| \leq z\right)\right| \to 0. \tag{6}$$

In terms of the linear contrasts $\lambda'\hat{\mathcal{Z}}_{(i)}$ we specifically care about $\lambda = [\mathbf{0}'_{k_\delta}, 1]'$ in view of expansion (5), hence we only require a pointwise limit theory vis-à-vis λ. The approximation does not require standardized $\lambda'\hat{\mathcal{Z}}_{(i)}$ and $\boldsymbol{Z}_{(i)}(\lambda)$ because we assume $\lambda'\mathcal{H}^{-1}_{(i)}\lambda$ lies in a compact subset of $(0,\infty)$ asymptotically uniformly in i and λ. We work in the now seminal setting of Chernozhukov, Chetverikov, and Kato (2013), cf. Chernozhukov, Chetverikov, and Kato (2017).

Theorem 2.4. *Let H_0 and Assumption 1 hold. Then $\rho_n(\lambda) = O(1/g(n)) \to 0$ for any slowly varying $g(n) \to \infty$, where by the (4) covariate cases (i)-(iii) $\ln(k_{\theta,n}) = o(n^{1/7})$, and (iv) $k_{\theta,n} = o(n^{p/2}/[g(n)\ln^2(n)]^p)$ where $p \geq 4$.*

Remark 2.4. In their post-estimation, DBL max-test, Dezeure, Bühlmann, and Zhang (2017) consider only covariate boundedness (i). They implicitly require slowly varying $g(n) \to \infty$ in their similar Gaussian approximation theory. Boundedness, however, need not hold in practice, and therefore does not fully describe the scope of required bounds on $k_{\theta,n}$. Although we also achieve $\ln(k_{\theta,n}) = o(n^{1/7})$ under sub-Gaussian and sub-exponential cases, under $\mathcal{L}_p$-boundedness the range reduces to $k_{\theta,n} = o(n^{p/2}/[g(n)\ln^2(n)]^p)$ where slowly varying $g(n)$ guides convergence of the Kolmogorov distance $\rho_n(\lambda)$. By construction, therefore, $k_{\theta,n} = o(n^{p/2-\iota})$ for infinitessimal $\iota > 0$.

Remark 2.5. Imposing slow variation $g(n) \to \infty$ optimizes the upper bound on $k_{\theta,n}$ under (iv). If a faster rate of convergence $\rho_n(\lambda) \to 0$ is desired, this will both reduce, and complicate solving for, an upper bound on $k_{\theta,n}$.

Remark 2.6. $k_{\theta,n}$ cannot have an exponential upper bound if higher moments do not exist. In this case, e.g., at least $k_{\theta,n} = o\left(n^2/[g(n)\ln^2(n)]^4\right)$ under $\mathcal{L}_4$-boundedness. If all moments exist, however, but a moment generating function near zero does not exist (e.g. log-normal: else we revert to cases (ii) or (iii)), then we only need $k_{\theta,n} = o\left(n^{\varphi}\right)$ for arbitrarily large $\varphi > 0$.

We now have the main result of this section. The statistic $\mathcal{T}_n = \max_{1\leq i\leq k_{\theta,n}} |\sqrt{n}\mathcal{W}_{n,i}\hat{\theta}_i|$ can be well approximated by the max-Gaussian process $\{\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_i \boldsymbol{Z}_{(i)}([\mathbf{0}'_{k_\delta}, 1])|\}_{n\in\mathbb{N}}$. The bounds on $k_{\theta,n}$ from the Gaussian approximation Theorem 2.4 dominate the asymptotic approximation Lemma 2.2 bounds except under $\mathcal{L}_p$-boundedness for finite p.

Theorem 2.5. *Let Assumption 1 hold, and assume for non-stochastic* $\mathcal{W}_i \in (0,\infty)$, $\max_{i\in\mathbb{N}} \mathcal{W}_i < \infty$, *that* $\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_{n,i} - \mathcal{W}_i| = O_p(\sqrt{\ln(k_{\theta,n})\mathcal{M}_n/n})$. *Assume by the (4) covariate cases* (i)-(iii) $\ln(k_{\theta,n}) = o(n^{1/7})$, *and* (iv) $k_{\theta,n} = o((n/\ln(n)^2)^{p/8})$ *where* $p \geq 4$.

a. Under H_0, *for the Gaussian random variables* $\boldsymbol{Z}_{(i)}(\lambda) \sim N(0, E[\epsilon_t^2]\lambda'\mathcal{H}_{(i)}^{-1}\lambda)$ *in (6),*

$$\left| P\left(\max_{1\leq i\leq k_{\theta,n}} \left|\sqrt{n}\mathcal{W}_{n,i}\hat{\theta}_i\right| \leq z\right) - P\left(\max_{1\leq i\leq k_{\theta,n}} \left|\mathcal{W}_i \boldsymbol{Z}_{(i)}([\mathbf{0}'_{k_\delta}, 1])\right| \leq z\right)\right| \to 0 \ \forall z \geq 0. \tag{7}$$

Furthermore, $\max_{1\leq i\leq k_{\theta,n}} |\sqrt{n}\mathcal{W}_{n,i}\hat{\theta}_i| \xrightarrow{d} \max_{i\in\mathbb{N}} |\mathcal{W}_i \boldsymbol{Z}_{(i)}([\mathbf{0}'_{k_\delta}, 1])|$.

b. Under H_1, $\max_{1\leq i\leq k_{\theta,n}} |\sqrt{n}\mathcal{W}_{n,i}\hat{\theta}_i| \to \infty$ *for any monotonic sequence of positive integers* $\{k_{\theta,n}\}$, $k_{\theta,n} \to \infty$.

Remark 2.7. We can attempt to characterize the max-statistic $\mathcal{T}_n$ null limit sequence $\{\max_{1\leq i\leq k} \left|\mathcal{W}_i \boldsymbol{Z}_{(i)}([\mathbf{0}'_{k_\delta}, 1])\right|\}_{k\in\mathbb{N}}$ after re-scaling. In Appendix D of the SM we present a simplified environment such that under H_0 for some positive non-random $b_n \sim \sqrt{2\ln(k_{\theta,n})}$, we have $\sqrt{2\ln(k_{\theta,n})/E[\epsilon_t^2]}(\mathcal{T}_n - b_n) \xrightarrow{d} \mathfrak{G}$, a Gumbel law. The next section, however, delivers a wild bootstrapped p-value approximation that allows for nuisance parameters and arbitrary covariate dependence, and naturally affords better small sample inference, *without* requiring knowledge of the limit law of a sequence of standardized $\mathcal{T}_n$.

3 Parametric Bootstrap

We utilize a variant of the wild (multiplier) bootstrap for inference (see Bose, 1988; Liu, 1988; Gonçalves and Kilian, 2004). Define the restricted estimator $\hat{\beta}^{(0)} \equiv [\hat{\delta}^{(0)\prime}, \mathbf{0}_{k_{\theta,n}}^{\prime}]^{\prime}$ under $H_0 : \theta_0 = \mathbf{0}_{k_{\theta,n}}$, where $\hat{\delta}^{(0)}$ minimizes the restricted least squares criterion $\sum_{t=1}^{n}(y_t - \delta^{\prime} x_{\delta,t})^2$. Define residuals $\epsilon_{n,t}^{(0)} \equiv y_t - \hat{\delta}^{(0)\prime} x_{\delta,t}$, draw iid $\{\eta_t\}_{t=1}^{n}$ from $N(0,1)$, and generate $y_{n,t}^{*} \equiv \hat{\delta}^{(0)\prime} x_{\delta,t} + \epsilon_{n,t}^{(0)} \eta_t$. Now construct regression models $y_{n,t}^{*} = \beta_{(i)}^{\prime} x_{(i),t} + v_{n,(i),t}$ for each $i = 1, ..., k_{\theta,n}$, and let $\hat{\tilde{\beta}}_{(i)} = [\hat{\tilde{\delta}}_{(i)}^{\prime}, \hat{\tilde{\theta}}_i]^{\prime}$ be the least squares estimator of $\beta_{(i)}$. The bootstrapped test statistic is

$$\tilde{\mathcal{T}}_n \equiv \max_{1 \leq i \leq k_{\theta,n}} \left| \sqrt{n} \mathcal{W}_{n,i} \hat{\tilde{\theta}}_i \right|. \tag{8}$$

Repeat the above steps $\mathcal{R}$ times, each time drawing a new iid sequence $\{\eta_{j,t}\}_{t=1}^{n}$, $j = 1 ... \mathcal{R}$. This results in a sequence of bootstrapped estimators $\{\hat{\tilde{\theta}}_j\}_{j=1}^{\mathcal{R}}$ and test statistics $\{\tilde{\mathcal{T}}_{n,j}\}_{j=1}^{\mathcal{R}}$ that are iid conditional on the sample $\{x_t, y_t\}_{t=1}^{n}$. The approximate p-value is

$$\tilde{p}_{n,\mathcal{R}} \equiv \frac{1}{\mathcal{R}} \sum_{j=1}^{\mathcal{R}} I(\tilde{\mathcal{T}}_{n,j} > \mathcal{T}_n). \tag{9}$$

The above algorithm varies from Gonçalves and Kilian (2004) since they operate on the *unrestricted* estimator in order to generate residuals. We impose the null when we estimate β_0, thus reducing dimensionality without affecting asymptotics under either hypothesis. Further, in view of the Lemma 2.2 asymptotic expansion, the parametric wild bootstrap is asymptotically equivalent to a wild (multiplier) bootstrap applied to the first order term $\hat{\mathcal{Z}}_{(i)}$. Simply draw iid $\{\eta_t\}_{t=1}^{n}$ from $N(0,1)$, compute $\hat{\tilde{\mathcal{Z}}}_{(i)} \equiv \sqrt{n} \widehat{\mathcal{H}}_{(i)}^{-1} \hat{\tilde{\mathcal{G}}}_{(i)}$ where $\hat{\tilde{\mathcal{G}}}_{(i)} \equiv -1/n \sum_{t=1}^{n} \eta_t (y_t -$

$\hat{\beta}'_{(i)}x_{(i),t})x_{(i),t}$, and then $\tilde{\mathcal{T}}_n \equiv \max_{1\leq i\leq k_{\theta,n}} |\sqrt{n}\mathcal{W}_{n,i}\widehat{\mathcal{H}}_{(i)}^{-1}\widehat{\tilde{\mathcal{G}}}_{(i)}|$.

Now write $\beta^{(0)} = [\delta^{(0)\prime}, \mathbf{0}'_{k_{\theta,n}}]'$ and $\beta^{(0)}_{(i)} = [\delta^{(0)\prime}, 0]'$, where $\delta^{(0)}$ minimizes $E[(y_t - \delta' x_{\delta,t})^2]$ on compact $\mathcal{D} \subset \mathbb{R}^{k_\delta}$. Define the normalized bootstrapped gradient under the null:

$$\tilde{\mathcal{Z}}^{(0)}_{(i)} \equiv \left(E\left[x_{\delta,t}x'_{\delta,t}\right]\right)^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \eta_t \left(y_t - \delta^{(0)\prime}x_{\delta,t}\right) x_{\delta,t}. \tag{10}$$

We have the following bootstrap expansion. Recall $\mathcal{M}_n \equiv E[\max_{1\leq i\leq k_{\theta,n}} |x_{(i),t}|^4]$.

The following Lemmas 3.1 and 3.2 and Theorem 3.3 offer improved bounds on $k_{\theta,n}$ due to the iid Gaussian multiplier term, and vagaries of the proofs of expansion and Gaussian approximation Lemmas 3.1 and 3.2.

Lemma 3.1. *Let Assumption 1 hold, and let the weights satisfy* $\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_{n,i} - \mathcal{W}_i| = O_p(\sqrt{\ln(k_{\theta,n})\mathcal{M}_n/n})$ *for non-stochastic* $\mathcal{W}_i \in (0,\infty)$. *Then:*

$$\left|\max_{1\leq i\leq k_{\theta,n}} \left|\sqrt{n}\mathcal{W}_{n,i}\widehat{\tilde{\theta}}_i\right| - \max_{1\leq i\leq k_{\theta,n}} \left|\mathcal{W}_i[\mathbf{0}'_{k_\delta}, 1]\tilde{\mathcal{Z}}^{(0)}_{(i)}\right|\right| = O_p\left(\ln(k_{\theta,n})\mathcal{M}_n^{3/4}/\sqrt{n}\right).$$

Hence the $O_p(\cdot)$ *term is* $o_p(1)$ *if by the (4) covariate cases: (i)* $\ln(k_{\theta,n}) = o(n^{1/2})$*; (ii)* $\ln(k_{\theta,n}) = o(n^{4/11})$*; (iii)* $\ln(k_{\theta,n}) = o(n^{2/7})$*; and (iv)* $k_{\theta,n} = o((n/[\ln(n)]^2)^{p/6})$, $p \geq 4$.

Remark 3.1. The upper bound $O_p(\ln(k_{\theta,n})\mathcal{M}_n^{3/4}/\sqrt{n})$ is smaller than the corresponding bound for $\hat{\theta}_i$ by a factor of $\mathcal{M}_n^{1/4}$ (hence convergence to zero is faster), ultimately due to the Gaussian multiplier. This in turn leads to greater upper bounds on $k_{\theta,n}$ by case compared to Lemma 2.2.

In order to characterize a conditional high dimensional Gaussian approximation for $\tilde{\mathcal{Z}}^{(0)}_{(i)}$, define $\bar{\mathcal{H}}^{(0)}_{(i)} \equiv E[x_{\delta,t}x'_{\delta,t}]$ and

$$\tilde{\sigma}^2_{(i)}(\lambda) \equiv \lambda' \left(\bar{\mathcal{H}}^{(0)}_{(i)}\right)^{-1} E\left[\left(y_t - \delta^{(0)\prime}x_{\delta,t}\right)^2 x_{\delta,t}x'_{\delta,t}\right] \left(\bar{\mathcal{H}}^{(0)}_{(i)}\right)^{-1} \lambda. \tag{11}$$

Notice $\tilde{\sigma}^2_{(i)}(\lambda) = E[\epsilon_t^2]\lambda'\mathcal{H}_{(i)}^{-1}\lambda$ under H_0 follows from Theorem 2.1. As usual, we only care about $\lambda = [\mathbf{0}'_{k_\delta}, 1]'$. Let $\Rightarrow^p$ denote *weak convergence in probability* (Giné and Zinn, 1990: Section 3), a useful notion for multiplier bootstrap asymptotics. We have the following conditional multiplier bootstrap central limit theorem.

Lemma 3.2. *Let Assumption 1 hold. Let* $\{\tilde{\boldsymbol{Z}}_{(i)}(\lambda)\}_{i\in\mathbb{N}}$*,* $\tilde{\boldsymbol{Z}}_{(i)}(\lambda) \sim N(0, \tilde{\sigma}^2_{(i)}(\lambda))$*, be an independent copy of the Theorem 2.5.a null distribution process* $\{\boldsymbol{Z}_{(i)}(\lambda)\}_{i\in\mathbb{N}}$*, that is independent of the asymptotic draw* $\{x_t, y_t\}_{t=1}^{\infty}$*. Let* $\{k_{\theta,n}\}$ *satisfy by the (4) covariate cases (i)-(iii)* $\ln(k_{\theta,n}) = o(n^{1/3})$*, and (iv)* $k_{\theta,n} = o((n/\ln(n))^{p/4})$*. Then* $\sup_{z\geq 0} |P(\max_{1\leq i\leq k_{\theta,n}} |\lambda'\tilde{\mathcal{Z}}^{(0)}_{(i)}| \leq z|\mathfrak{S}_n) - P(\max_{1\leq i\leq k_{\theta,n}} |\tilde{\boldsymbol{Z}}_{(i)}(\lambda)| \leq z)| \xrightarrow{p} 0$*. Furthermore* $\max_{1\leq i\leq k_{\theta,n}} |[\mathbf{0}'_{k_\delta}, 1]\tilde{\mathcal{Z}}^{(0)}_{(i)}| \Rightarrow^p \max_{i\in\mathbb{N}} |\tilde{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta}, 1])|$.

Bootstrap asymptotic expansion Lemma 3.1, and bootstrap Gaussian approximation Lemma 3.2, coupled with $|\mathcal{W}_{n,i} - \mathcal{W}_i|_\infty = O_p(\sqrt{\ln(k_{\theta,n})\,\mathcal{M}_n/n})$ prove the following claim. The bounds on $k_{\theta,n}$ are merely the least yielded from the two supporting lemmas.

Theorem 3.3. *Let Assumption 1 hold, and for non-stochastic* $\mathcal{W}_i \in (0,\infty)$*,* $\max_{i\in\mathbb{N}} \mathcal{W}_i < \infty$*, assume* $\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_{n,i} - \mathcal{W}_i\,| = O_p(\sqrt{\ln(k_{\theta,n})\,\mathcal{M}_n/n})$*. Let* $\{k_{\theta,n}\}$ *satisfy by the (4) covariate cases (i)* $\ln(k_{\theta,n}) = o(n^{1/3})$*; (ii)* $\ln(k_{\theta,n}) = o(n^{1/3})$*; (iii)* $\ln(k_{\theta,n}) = o(n^{2/7})$*; and (iv)* $k_{\theta,n} = o((n/[\ln(n)]^2)^{p/6})$*,* $p \geq 4$*. Let* $\{\tilde{\boldsymbol{Z}}_{(i)}(\lambda)\}_{i\in\mathbb{N}}$*,* $\tilde{\boldsymbol{Z}}_{(i)}(\lambda) \sim N(0, \tilde{\sigma}^2_{(i)}(\lambda))$*, be an independent copy of the Theorem 2.5.a null distribution process* $\{\boldsymbol{Z}_{(i)}(\lambda)\}_{i\in\mathbb{N}}$*, that is independent of the asymptotic draw* $\{x_t, y_t\}_{t=1}^{\infty}$*. Then* $\max_{1\leq i\leq k_{\theta,n}} |\sqrt{n}\mathcal{W}_{n,i}\hat{\tilde{\theta}}_i| \Rightarrow^p \max_{i\in\mathbb{N}} |\mathcal{W}_i\tilde{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta}, 1])|$.

We now have the main result of this section: the approximate p-value $\tilde{p}_{n,\mathcal{R}_n}$ promotes a correctly sized and consistent test asymptotically. Notice that while Lemmas 3.1 and 3.2 offer improvements on the upper bound for $k_{\theta,n}$, here we have the reduced bounds set by Theorem 2.5 because $\tilde{p}_{n,\mathcal{R}_n}$ is computed from the bootstrapped $\hat{\tilde{\theta}}_i$ *and* original $\hat{\theta}_i$.

Theorem 3.4. *Let Assumption 1 hold, and for non-stochastic* $\mathcal{W}_i \in (0,\infty)$*,* $\max_{i\in\mathbb{N}} \mathcal{W}_i < \infty$*, assume* $\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_{n,i} - \mathcal{W}_i\,| = O_p(\sqrt{\ln(k_{\theta,n})\,\mathcal{M}_n/n})$*. Let* $\{k_{\theta,n}\}$ *satisfy by the (4) covariate cases (i)-(iii)* $\ln(k_{\theta,n}) = o(n^{1/7})$*, and (iv)* $k_{\theta,n} = o((n/\ln(n)^2)^{p/8})$ *where* $p \geq 4$.

a. Under H_0*,* $P(\tilde{p}_{n,\mathcal{R}_n} < \alpha) \to \alpha$*; and b. Under* H_1*,* $P(\tilde{p}_{n,\mathcal{R}_n} < \alpha) \to 1$*.*

Remark 3.2. The test is consistent against a global alternative despite the fact that $\mathcal{T}_n$ increases in magnitude in n, even under H_0. The proof reveals $\tilde{p}_{n,\mathcal{R}_n}$ is asymptotically in probability equivalent to

$$P\left(\max_{1\leq i\leq k_{\theta,n}} \left|\mathcal{W}_i\tilde{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])\right| > \max_{1\leq i\leq k_{\theta,n}} \left|\sqrt{n}\mathcal{W}_{n,i}\left(\hat{\theta}_i - \theta_{0,i}\right) + \sqrt{n}\mathcal{W}_i\theta_{0,i} + o_p(1)\right|\right),$$

where $\{\tilde{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])\}$ is an independent copy of zero-mean Gaussian $\{\boldsymbol{Z}_{(i)}(\lambda)\}_{i\in\mathbb{N}}$. The max-statistic is therefore $O_p(\sqrt{\ln(k_{\theta,n})\mathcal{M}_n} + \sqrt{n})$ under H_1, cf. Lemma A.1 in the appendix. Conversely, standard Gaussian concentration bounds yield $\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_i\tilde{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])| = O_p(\sqrt{\ln(k_{\theta,n})})$, cf. (A.4) in the appendix. This yields consistency $\tilde{p}_{n,\mathcal{R}_n} \xrightarrow{p} 0$.

Finally, the bootstrapped test has non-trivial power against a sequence of local alternatives $H_1^L : \theta_0 = c\sqrt{\ln(k_{\theta,n})\mathcal{M}_n/n}$ where $c = [c_i]_{i=1}^{k_{\theta,n}}$. The local-to-null rate follows from global max-asymptotics Lemma A.1.c. The rate is generally less than $\sqrt{n}$ because (i) we need to deal with $\max_{1\leq i\leq k_{\theta,n}} ||\widehat{\mathcal{H}}_{(i)}^{-1} - \mathcal{H}_{(i)}^{-1}||$ and hence the fourth moment envelope $\mathcal{M}_n$, cf. Remark 2.2, where generally $\mathcal{M}_n \to \infty$ is possible (e.g. for unbounded covariates); and (ii) the test statistic under the null is, $awp1$, the maximum of an increasing sequence of Gaussian random variables $\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_i\tilde{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])|$, and $\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_i\tilde{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])| = O_p(\sqrt{\ln(k_{\theta,n})})$ as discussed above

Theorem 3.5. *Under the conditions of Theorem 3.4,* $\lim_{n\to\infty} P(\tilde{p}_{n,\mathcal{R}_n} < \alpha) \nearrow 1$ *under* H_1^L *monotonically as the maximum local drift* $\max_{i\in\mathbb{N}} |c_i| \nearrow \infty$*.*

Remark 3.3. Asymptotic local power therefore does not depend on the degree, if any, of sparsity. By comparison, for example, Zhong, Chen, and Xu (2013) consider a high dimensional mean model $y_i = \theta_0 + \epsilon_i$ and test $H_0 : \theta_0 = \mathbf{0}_p$, where $y_i \in \mathbb{R}^p$. Under their alternative $p^{1-\xi}$ elements $\theta_{0,i} \neq 0$ for some $\xi \in (1/2, 1)$, where $\theta_{0,i} = \sqrt{r\ln(p)/n}$ for some $r > 0$, and by assumption $\ln(p) = o(n^{1/3})$, hence $\theta_{0,i} = o(1/n^{2/3})$. See also Arias-Castro, Candès, and Plan (2011) and their references. Our max-test is consistent against local deviations $\theta_0 = c(\ln(k_{\theta,n})/n)^{\zeta}$ for any $\zeta \in [0, 1/2)$ and $c \neq \mathbf{0}_{k_{\theta,n}}$, and has non-trivial power when $\zeta = 1/2$ and $c \neq \mathbf{0}_{k_{\theta,n}}$. Further, we allow for any positive number of elements $\theta_{0,i} \neq 0$ under H_1 or H_1^L (i.e. $1 \leq \sum_{i=0}^{k_{\theta,n}} I(\theta_{0,i} \neq 0) \leq k_{\theta,n}$). Thus, in the simple setting of Zhong, Chen, and Xu (2013), alternatives can be closer to the null than allowed here in a max-test setting, but sparsity is enforced. Here, any degree of (non)sparsity is permitted under H_1, (as long as some $\theta_{0,i} \neq 0$) or H_1^L (provided $c_i \neq 0$ for some i).

4 Monte Carlo Experiments

We use 1000 independently drawn samples of sizes $n \in \{100, 250, 500\}$. We perform two max-tests: one with flat $\mathcal{W}_{n,i} = 1$, and the other $\mathcal{W}_{n,i}$ is equal to the inverted least squares standard error of $\hat{\theta}_{n,i}$, resulting in the *max-test* and *max-t-test*. Other tests are discussed below. Unless otherwise noted, all bootstrapped p-values are based on $\mathcal{R} = 1000$ independently drawn samples from $N(0,1)$. Significance levels are $\alpha \in \{.01, .05, .10\}$.

The following presents our benchmark design. Robustness checks are performed in Appendix E of the supplemental material.

4.1 Design and Tests of Zero Restrictions

4.1.1 Set-Up

The DGP is $y_t = \delta_0' x_{\delta,t} + \theta_0' x_{\theta,t} + \epsilon_t$ where ϵ_t is iid standard normal. The total number of covariates is $k \equiv k_\delta + k_{\theta,n}$. We consider two opposing nuisance parameter cases $k_\delta \in \{0, 10\}$, and for $x_{\theta,t}$ we set $k_{\theta,n} \in \{35, k_1(n), k_2(n)\}$. We use $k_1(n) \equiv [\exp\{3.2 n^{1/7-\iota}\}]$ with $\iota = 10^{-10}$, which is valid asymptotically when x_t is bounded (cf. Theorem 3.4). We also use $k_2(n) =$

$[.02n^2]$, which is valid asymptotically in any case here.[4] Thus in this experiment $k_{\theta,100} \in \{35, 200, 482\}$, $k_{\theta,250} \in \{35, 1144, 1250\}$ and $k_{\theta,500} \in \{35, 2381, 5000\}$.

We considered nine total covariate cases $[x_{\delta,t}, x_{\theta,t}]$ (three bound cases and three dependency cases). In three broad scenarios $x_{i,t}$ are random draws from a truncated standard normal distribution, denoted $\mathcal{TN}(0,1,U)$, such that $-U \leq x_{i,t} \leq U$ $a.s.$, with bounds $U_1 = 2.5$, $U_2 = 10^{10}$, or $U_3 = \infty$ (i.e. $x_{i,t} \sim \mathcal{N}(0,1)$). Only upper bound U_1 has an impact since under U_2 the bound was never surpassed in our experiment. The theory under boundedness, however, allows for any $k_{\theta,n}$ provided $\ln(k_{\theta,n}) = o(n^{1/7})$.

Under these three scenarios we then consider three dependence types for $[x_{\delta,t}, x_{\theta,t}]$: first, $[x_{\delta,t}, x_{\theta,t}]$ are serially and mutually independent standard normals; second, they are block-wise dependent normals ($x_{\delta,t}$ and $x_{\theta,t}$ are mutually independent); and third, they are within and across block dependent normals.

Type three covariates are drawn as follows. Combine $x_t \equiv [x'_{\delta,t}, x'_{\theta,t}]' \in \mathbb{R}^k$, and let (w_t, v_t) be mutually independent draws $\mathcal{TN}(0, I_k)$. The regressors are $x_t = Aw_t + v_t$. We randomly draw each element of $A \in \mathbb{R}^k$ from a uniform distribution on $[-1,1]$. If A does not have full column rank then we add a randomly drawn ι from $[0,1]$ to each diagonal component (in this study every draw A had full column rank).

Not surprisingly all tests perform better when signals are less entangled. Since our U_2-bounded and U_3-unbounded designs lead to identical results, we therefore only report results for U_1-bounded and U_3-unbounded x_t under dependence type three (within and across block dependent).

We fix $\delta = \mathbf{1}_{k_\delta}$, a $k_\delta \times 1$ vector of ones. The benchmark models are as follows. Under the null $\theta_0 = [\theta_{0,i}]_{i=1}^{k_{\theta,n}} = \mathbf{0}_{k_{\theta,n}}$. The alternatives are Alt(i) $\theta_{0,1} = .0015$ with $[\theta_{0,i}]_{i=2}^{k_{\theta,n}} = \mathbf{0}_{k_{\theta,n}-1}$; Alt($ii$) $\theta_{0,i} = .0002i/k_{\theta,n}$ for $i = 1, ..., k_{\theta,n}$; and Alt(iii) $\theta_{0,i} = .00015$ for $i = 1, ..., 10$. See Table 1 for reference. Thus, under Alt(i) only one element deviates from, but is close to, zero. Alts

[4]In our design x_t is bounded or Gaussian and is therefore $\mathcal{L}_p$-bounded for any $p > 0$.

(ii) and (iii) have smaller values, but for several or all $\theta_{0,i}$.

Table 1: Alternative Models

$$\begin{aligned} \text{Alt}(i): \quad & \theta_{0,1} = .0015 \text{ and } [\theta_{0,i}]_{i=2}^{k_{\theta,n}} = 0 \\ \text{Alt}(ii): \quad & \theta_{0,i} = .0002 \times i/k_{\theta,n} \text{ for } i = 1, ..., k_{\theta,n} \\ \text{Alt}(iii): \quad & \theta_{0,i} = .00015 \text{ for } i = 1, ..., 10 \end{aligned}$$

4.1.2 Parsimonious Max-Tests

We estimate $k_{\theta,n}$ parsimonious models $y_t = \delta_{(i)}^{*\prime} x_{\delta,t} + \theta_i^* x_{\theta,i,t} + v_{(i),t}$ by least squares. Denote by $\mathcal{T}_n$ the resulting max-test or max-t-test statistic. The bootstrapped test statistic $\tilde{\mathcal{T}}_n$ and p-value $\tilde{p}_{n,\mathcal{R}}$ are computed as in (8) and (9). We reject H_0 when $\tilde{p}_{n,\mathcal{R}} < \alpha$.

4.1.3 De-biased Lasso Max-Tests

We follow Dezeure, Bühlmann, and Zhang (2017), cf. Zhang and Zhang (2014), and estimate (1) by DBL, using 5-fold cross-validation in order to select the tuning parameter. Dezeure, Bühlmann, and Zhang (2017) propose a (post-estimation) max-t-test for a parameter subset, and various bootstrap techniques. We perform both max- and max-t-tests for direct comparisons with the parsimonious max-tests. The max-t-test uses the standard error in Dezeure, Bühlmann, and Zhang (2017, eq. (4)), de-biasing is performed node-wise only for θ, and p-values are approximated by parametric wild bootstrap as above.

Computation time for bootstrapped DBL is prohibitive. In Footnote 3 we saw our method is up to $1600x$ faster when $n = 100$ and $35,000x$ faster when $n = 250$, based on using 128 workers in a parallel processing environment. The speed difference, however, clouds just how slow bootstrapped DBL is: just one p-value under $n = 100$ with $k_{\theta,n} = 480$ took over 2 hours, and over 50 hours when $n = 250$ and $k_{\theta,n} = 1144$. Scaling up to 1000 samples is obviously not currently attractive. We therefore ran a limited experiment for DBL using only $n = 100$ with 250 samples, and 250 bootstrap samples.

4.1.4 Wald Tests

If $k_\delta + k_{\theta,n} < n$ then we also perform asymptotic and bootstrapped Wald tests, and asymptotic and bootstrapped normalized Wald tests. This will highlight in non-high dimensional settings the advantages of a max-test over a Wald test. We estimate $y_t = \delta_0' x_{\delta,t} + \theta_0' x_{\theta,t} + \epsilon_t$ by least squares. The asymptotic Wald test, based on the $\chi^2(k_{\theta,n})$ distribution, leads to large empirical size distortions and is therefore not reported here. The bootstrapped Wald test is based on a parametric wild bootstrap.

The normalized Wald statistic is $\mathcal{W}_n^s \equiv (\mathcal{W}_n - k_{\theta,n})/\sqrt{2k_{\theta,n}}$. Under the null $\mathcal{W}_n^s \xrightarrow{d} N(0,1)$ as $k_{\theta,n} \to \infty$, and as long as $k_{\theta,n}/n \to 0$ then $\mathcal{W}_n^s \xrightarrow{p} \infty$ under H_1. This asymptotic one-sided test rejects the null when $\mathcal{W}_n^s > Z_\alpha$, where Z_α is the standard normal upper tail α-level critical value. This test yields highly distorted empirical sizes and is therefore not reported. The bootstrapped test uses $\mathcal{W}_{n,i}^{s*} \equiv (\mathcal{W}_{n,i}^* - k_{\theta,n})/\sqrt{2k_{\theta,n}}$ with p-value approximation $p_n = 1/\mathcal{R} \sum_{i=1}^{\mathcal{R}} I(\mathcal{W}_n^s > \mathcal{W}_{n,i}^{s*})$. Trivially $\mathcal{W}_n^s > \mathcal{W}_{n,i}^{s*}$ *if and only if* $\mathcal{W}_{n,i}^* > \mathcal{W}_n$, hence we only discuss the bootstrapped Wald test.

4.2 Simulation Results

4.2.1 Benchmark Results

We report rejection frequencies in Tables 2 and 3 when nuisance terms $k_\delta = 0$, with unbounded x_t. See Appendix F of SM for complete simulation results (bounded and unbounded x_t; $k_\delta \geq 0$).

Empirical Size Let *p*-max and *dbl*-max denote parsimonious and DBL max-tests. Test results under the three covariate cases are similar, although all tests perform slightly less well under correlated regressors within and across blocks (case three). We therefore only report and comment on the latter results.

The p-max tests typically lead to qualitatively similar results with empirical size close to nominal size (see Table 2). The tests perform roughly the same whether covariates are bounded or not. In the presence of nuisance parameters $k_{\theta,n} = 10$ the tests are slightly over-sized when $n = 100$ with improvements as n increases (the tests are comparable when $n \geq 250$).

The *dbl*-max tests yield competitive empirical size when $k_\delta = 0$, although the *dbl*-max test is generally over-sized compared to the p-max tests, especially when $k_{\theta,n}$ is large. Furthermore, the *dbl*-max t-test tends to be highly under-sized when there are nuisance parameters ($k_\delta = 10$). The latter naturally implies comparatively low power, discussed below. Overall, across cases the p-tests yield better empirical size than the *dbl*-max tests.

In the case $k_\delta + k_{\theta,n} < n$, the asymptotic Wald test and asymptotic normalized Wald test are generally severely over-sized due to the magnitude of $k_\delta + k_{\theta,n}$ (the latter test is not shown). The bootstrapped Wald test is strongly under-sized when n is small, but also when $k_\delta = 10$ and/or when $k_{\theta,n}$ is large.

Empirical Power See Tables 2 and 3. The bootstrapped Wald test in a low dimension setting yields lower empirical power than the p-max tests in every case. The asymptotic Wald test yields size corrected power that is nearly zero. We now focus on the max tests.

Alt(i) Only one parameter exhibits a small deviation from the null: $\theta_{0,1} = .0015$. The difficulty in detecting such a deviation is apparent at $n = 100$ when nuisance parameters are present and/or when $k_{\theta,n}$ is small. The latter suggests a noise-signal effect: when $k_{\theta,n} >> n$ the impact of the $k_\delta = 10$ nuisance parameters is negligible, but the impact can be large under any alternative studied here when $k_{\theta,n}$ is small. For example, the p-max-tests yield (nearly) 100% empirical power for sample sizes (100) 250 and 500 when $k_\delta = 0$, or when $k_\delta = 10$ and $k_{\theta,n} >> n$.

If $n = 100$, $k_{\theta,n} = 200$, $k_\delta = 10$ and covariates are bounded then power drops to $(.941, .975, .988)$ for p-max-t at sizes $(.01, .05, .10)$, and when covariates are unbounded power is comparable at $(.974, .993, .997)$. At $n = 250$ with $k_{\theta,n} >> n$ power rises to 100% for bounded and unbounded covariates. Overall, p-max-test performance is not systematically sensitive to whether the covariates are bounded.

The *dbl*-max-tests are generally dominated by the p-max tests, in many cases significantly so. Consider $n = 100$, $k_{\theta,n} = 200$, $k_\delta = 10$ with bounded covariates: empirical power is only $(.002, .002, .003)$ for the *dbl*-max-t-test, compared to $(.941, .975, .988)$ for the p-max-t-test. If covariates are unbounded and there are no nuisance parameters, then *dbl*-max-t-test power improves to $(.132, .188, .220)$, compared to p-max-t-test power $(1.00, 1.00, 1.00)$.

The *dbl*-max-test yields generally low power even when $k_\delta = 0$: under $n = 100$, $k_{\theta,n} = 482$ with unbounded covariates, *dbl*-max power is (036, .180, .332) compared to $(.804, .900, .904)$ for *dbl*-max-t. However, when $k_{\theta,n} = 200$ power for both tests is under 22% for any nominal size compared to 100% across sizes for p-max and p-max-t.

Considering *dbl*-max has sub-par power, in the remaining discussion we focus only on *dbl*-max-t.

Alt(ii) In this case all parameters deviate from zero monotonically between $(0, .0002]$. The p-max test yields 100% empirical power when $k_\delta = 0$, or when $k_\delta = 10$ and $k_{\theta,n} >> n$, irrespective of covariate (un)boundedness, again suggesting a signal/noise effect.

The *dbl*-max tests, by comparison, are incapable of detecting such a signal when nuisance parameters are present and $n = 100$, yielding lower rejection rates than nominal size. That power is trivial may be an artifact of the design, and the requirements for DBL: (i) We do not need to impose sparsity because our method does not require it. Yet (ii) DBL requires sparsity (e.g. Dezeure, Bühlmann, and Zhang, 2017, (B.2)): DBL is both picking up the strong nuisance parameter values *and* effectively setting $\hat{\theta}_i \approx 0$. And (iii) bias follows under sparsity failure, in this case resulting in rejection rates below size.

The *dbl*-max tests yield decent empirical power when nuisance parameters are not present, although consistently below, or far below, p-max tests. This applies whether covariates are bounded (as in Dezeure, Bühlmann, and Zhang, 2017) or unbounded.

Alt(iii) This setting is a cross-hatch of Alts (i) and (ii). Here we have smaller $\theta_{0,i} = .00015$ than Alt(i) for $i = 1, ..., 10$, for more components than Alt(i) but fewer than Alt(ii). As in previous cases p-max test power is (nearly) unity when $k_\delta = 0$, or when $k_\delta = 10$ and $k_{\theta,n} >> n$. In the presence of nuisance parameters and $n = 100$, power is noticeably smaller: p-max test power is $(.039, .174, .342)$ and $(.058, .253, .450)$ when $k_{\theta,n} = 200$ and 482 respectively. Those values are near or exactly unity once $n = 250$.

The *dbl*-max tests, by comparison, yield trivial power with $n = 100$, under either covariate case, and for any $k_\delta \geq 0$. The signal appears to be too weak to detect when $n = 100$, in particular given the shrinkage nature of DBL.

5 Conclusion

We present a class of tests for a high dimensional parameter in a regression setting with nuisance parameters. We focus on a linear model and least squares estimation, but the method and theory presented here are generalizable to a broad class of models and estimators. We show how the hypotheses can be identified by using many $(k_{\theta,n})$ potentially vastly lower dimension parameterizations depending on the existence of nuisance parameters, and we allow for up to $\ln(k_{\theta,n}) = o(n^{1/7})$ low dimension models, depending on covariate assumptions. We use a test statistic that is the maximum in absolute value of the weighted key estimated parameters across the many low dimension settings. The lower dimension helps improve estimation accuracy and allows us to sidestep sparsity requirements, while a max-statistic both alleviates the need for a multivariate normalization used in Wald and score statistics, and hones in the most informative (weighted) model component. Thus, we avoid inverting a potentially large dimension variance estimator that may be a poor proxy for the true sampling dispersion. Indeed, our asymptotic theory sidesteps traditional extreme value theory arguments, while we instead focus on an approximate p-value computed by parametric wild bootstrap.

In simulation experiments our max-tests generally have good or sharp size. They generally dominate bootstrapped debiased Lasso based max-tests in terms of power, and bootstrapped (and normalized) Wald test in terms of both size and power in low dimension cases.

Future work may (i) lean toward allowing for a broader class of semi-nonparametric models, including models with a high dimensional parameter θ_0 *and* infinite dimensional unknown function h and/or high dimensional nuisance term δ_0; (ii) verify key asymptotic theory arguments in a general time series setting; and (iii) expand McKeague and Qian's (2015) significant predictor method to a high dimensional setting.

A Appendix

Throughout $O_p(1)$ and $o_p(1)$ are not functions of model counter i. $\{k_{\theta,n}, k_n\}$ are monotonically increasing sequences of positive integers. Define parsimonious parameter spaces $\mathcal{B}_{(i)} \equiv \mathcal{D} \times \Theta_{(i)}$ where $\mathcal{D} \subset \mathbb{R}^{k_\delta}$ and $\Theta_{(i)} \subset \mathbb{R}$ are compact subsets, δ_0 is an interior point of $\mathcal{D}$, and 0 and $\theta_{0,i}$ are interior points of $\Theta_{(i)}$. Define compact $\mathcal{B} \equiv \mathcal{D} \times \Theta$ where $\Theta \equiv \times_{i=1}^{k_{\theta,n}} \Theta_i$. Recall $\widehat{\mathcal{H}}_{(i)} \equiv 1/n \sum_{t=1}^{n} x_{(i),t} x_{(i),t}'$ and $\mathcal{H}_{(i)} \equiv E[x_{(i),t} x_{(i),t}']$. Let $\underline{\lambda}_{(i),n}$ and $\underline{\lambda}_{(i)}$ denote the minimum eigenvalues of $\widehat{\mathcal{H}}_{(i)}$ and $\mathcal{H}_{(i)}$

A.1 Assumptions

Assumption 1.

a. (ϵ_t, x_t) *are iid over* t; $E[\epsilon_t] = 0$; $\underline{c} \leq E[\epsilon_t^2] \leq \bar{c}$ *and* $\underline{c} \leq E[x_{j,t}^2] \leq \bar{c}$ *for all* j *and some* $\underline{c}, \bar{c} \in (0, \infty)$ *that may differ for different variables;* $E[\epsilon_t^4] < \infty$ *and* $E[x_{j,t}^4] < \infty$ *for all* j; $P(E[\epsilon_t^2 | x_t] = \sigma^2) = 1$ *for finite* $\sigma^2 > 0$; *and* $\limsup_{n\to\infty} |\theta_0| < \infty$.

b. β_0 *uniquely minimizes* $E[(y_t - \beta' x_t)^2]$ *on* $\mathcal{B}$; $E[(y_t - \beta_{(i)}^{*\prime} x_{(i),t}) x_{(i),t}] = \mathbf{0}_{k_\delta+1}$ *for all* i *and unique* $\beta_{(i)}^*$ *in the interior of* $\mathcal{B}_{(i)}$.

c. $\liminf_{n\to\infty} \min_{1\leq i\leq k_{\theta,n}} \inf_{\lambda'\lambda=1} E[(\lambda' x_{(i),t})^2] > 0$ *and* $\liminf_{n\to\infty} \min_{1\leq i\leq k_{\theta,n}} \{\underline{\lambda}_{(i)}\} > 0$.

d. $\liminf_{n\to\infty} \inf_{\lambda'\lambda=1} \min_{i\in\mathbb{N}} \{\frac{1}{n} \sum_{t=1}^{n} (\lambda \mathcal{H}_{(i)}^{-1\prime} x_{(i),t})^2\} > 0$ *a.s.;* $\liminf_{n\to\infty} \min_{i\in\mathbb{N}} \{\underline{\lambda}_{(i),n}\} > 0$ *a.s.*

Remark A.1. (a)-(d) are standard for linear iid regression models, augmented to allow for high dimension. The existence of higher moments is not unusual for a high dimensional maximum limit theory (e.g. Bühlmann and Van De Geer, 2011; Chernozhukov, Chetverikov, and Kato, 2013; Dezeure, Bühlmann, and Zhang, 2017; Zhang and Wu, 2017). Indeed, in order to handle the Gaussian approximation error $\max_{1\leq i\leq k_{\theta,n}}|[\mathbf{0}_{k_\delta}', 1]\hat{\mathcal{R}}_i|$ in (3), an upper bound on feasible $k_{\theta,n}$ is generally linked to higher moments. See the discussion following Lemma 2.2.

Remark A.2. By the constructions of the $\mathcal{L}_2$-minimizers β_0 and $\beta_{(i)}^*$, (1) and the parsimonious versions may be misspecified in the senses of Sawa (1978) and White (1982).

Remark A.3. We use the eigenvalue bounds $\liminf_{n\to\infty}\min_{1\leq i\leq k_{\theta,n}}\{\underline{\lambda}_{(i),n}\} > 0$ $a.s.$ to ensure $\max_{1\leq i\leq k_{\theta,n}}||\widehat{\mathcal{H}}_{(i)}^{-1}|| = \max_{1\leq i\leq k_{\theta,n}}\{\underline{\lambda}_{(i),n}^{-1}\} = (\min_{1\leq i\leq k_{\theta,n}}\{\underline{\lambda}_{(i),n}^{-1}\})^{-1} = O_p(1)$ is well defined for each n and any $\{k_{\theta,n}\}$. The condition $|\theta_0| < \infty$ uniformly in n is only used in Lemma A.1 in order to bound the parsimonious model error moment $E[\max_{1\leq i\leq k_{\theta,n}}\left|v_{(i),t}\right|^4]$ under *either* hypothesis. Under the null, of course, $E[\max_{1\leq i\leq k_{\theta,n}}\left|v_{(i),t}\right|^4] = E|\epsilon_t|^4 < \infty$.

A.2 Proofs of Main Results

Write compactly $|\cdot|_\infty \equiv \max_{1\leq i\leq k_{\theta,n}}|\cdot|$ and $||\cdot||_\infty \equiv \max_{1\leq i\leq k_{\theta,n}}||\cdot||$.

Proof of Theorem 2.1. By supposition $E[(y_t - \beta' x_t)x_{i,t}] = 0$ $\forall i$ *if and only if* $\beta = \beta_0 = [\delta_0', \theta_0']'$. Therefore, if $\theta_0 = \mathbf{0}_{k_{\theta,n}}$ then $E[(y_t - \delta_0' x_{\delta,t})x_{i,t}] = 0$ $\forall i$ hence $E[(y_t - \delta_0' x_{\delta,t})x_{(i),t}] = \mathbf{0}_{k_\delta+1}$ for each $i = 1,...,k_{\theta,n}$. This instantly yields $\theta_i^* = 0$ for each i and $\delta^* = \delta_0$ by construction and uniqueness of $\beta_{(i)}^* = [\delta^{*\prime}, \theta_i^*]'$.

Conversely, if $\theta^* = \mathbf{0}_{k_{\theta,n}}$ then from the parsimonious risk it follows $E[(y_t - \delta_{(i)}^{*\prime} x_{\delta,t})x_{(i),t}] = \mathbf{0}_{k_\delta+1}$ for each i by identification of $\beta_{(i)}^* = [\delta^{*\prime}, \theta_i^*]'$. But this implies the sub-gradient $E[(y_t - \delta_{(i)}^{*\prime} x_{\delta,t})x_{\delta,t}] = \mathbf{0}_{k_\delta+1}$ for each i, hence the $\delta_{(i)}^{*\prime}s$ must be identical: there exists a unique $\delta^* \in \mathcal{D}$ such that $\delta_{(i)}^* = \delta^*$ for each i. But $E[(y_t - \delta^{*\prime} x_{\delta,t})x_{(i),t}] = \mathbf{0}_{k_\delta+1}$ $\forall i$ yields $E[(y_t - \delta^{*\prime} x_{\delta,t})x_{j,t}] = 0$ $\forall j$ by the definition of $x_{(i),t}$. Therefore $[\delta_0, \theta_0] = [\delta^*, \mathbf{0}_{k_{\theta,n}}]$ by identification of $[\delta_0, \theta_0]$. $\mathcal{QED}$.

Proof of Theorem 2.4. Define[5] $s_{(i),t} = s_{(i),t}(\lambda) \equiv \lambda'(E[x_{(i),t}x_{(i),t}'])^{-1}x_{(i),t}$; $\mathcal{M}_r \equiv ||\epsilon_t||_r \times \max_{1\leq i\leq k_{\theta,n}}\left\|s_{(i),t}\right\|_r$ for $r \in (2,4]$, where $\limsup_{n\to\infty}\mathcal{M}_r < \infty$ under Assumption 1; $\zeta_{(i),t} \equiv \epsilon_t s_{(i),t}/(||\epsilon_t||_2||s_{(i),t}||_2)$; define random variables $z_{(i),t} \sim N(0,1)$ independent over t; and for any $\gamma \in (0,1)$ define $u(\gamma) \equiv u_{\epsilon s}(\gamma) \vee u_z(\gamma)$ where

$$u_{\epsilon s}(\gamma) \equiv \inf\left\{u : P\left(\max_{1\leq i\leq k_{\theta,n}, 1\leq t\leq n}\left|\zeta_{(i),t}\right| > u\right) \leq \gamma\right\}$$
$$u_z(\gamma) \equiv \inf\left\{u : P\left(\max_{1\leq i\leq k_{\theta,n}, 1\leq t\leq n}\left|z_{(i),t}\right| > u\right) \leq \gamma\right\}.$$

[5] We suppress λ, $\lambda'\lambda = 1$, when confusion is avoided.

Write $\mathcal{K}_{3,4} \equiv \mathcal{M}_3^{3/4} \vee \mathcal{M}_4^{1/2}$. Under Assumption 1.a,c and H_0, Theorem 2.2 in Chernozhukov, Chetverikov, and Kato (2013) applies to $\epsilon_t s_{(i),t}$. Hence, for every $\gamma \in (0,1)$, and some finite constant $\mathcal{C} > 0$ that depends only on lower and upper bounds on $E[\epsilon_t^2]$ and $E[s_{j,t}^2]$,

$$\rho_n(\lambda) \leq \mathcal{C}\left[\frac{1}{n^{1/8}}\mathcal{K}_{3,4}\left(\ln\left(\frac{k_{\theta,n}n}{\gamma}\right)\right)^{7/8} + \frac{1}{n^{1/2}}\left(\ln\left(\frac{k_{\theta,n}n}{\gamma}\right)\right)^{3/2} u(\gamma) + \gamma\right]. \tag{A.1}$$

We need to bound $u(\gamma)$. Notice

$$\sup_{\lambda'\lambda=1} E\left[\max_{1\leq i\leq k_{\theta,n}} s_{(i),t}^4(\lambda)\right] \leq \left|\left(E[x_{(i),t}x_{(i),t}']\right)^{-1}\right|_\infty^4 E\left|x_{(i),t}\right|_\infty^4 = \left|\mathcal{H}_{(i)}^{-1}\right|_\infty^4 \mathcal{M}_n,$$

where $\mathcal{M}_n \equiv E|x_{(i),t}|_\infty^4$, and $|\mathcal{H}_{(i)}^{-1}|_\infty \in (0,\infty)$ by Assumption 1.c. Assumption 1.c ensures $\chi \equiv \liminf_{n\rightarrow\infty} \min_{1\leq i\leq k_{\theta,n}} \inf_{\lambda'\lambda=1} ||s_{(i),t}(\lambda)||_2 > 0$. Hence, by the Cauchy-Schwartz inequality:

$$\begin{aligned}\left(E\left[\max_{1\leq i\leq k_{\theta,n}} \zeta_{(i),t}^2\right]\right)^{1/2} &\leq \frac{\|\epsilon_t\|_4}{\|\epsilon_t\|_2}\left(E\left[\max_{1\leq i\leq k_{\theta,n}} \frac{s_{(i),t}^4}{\|s_{(i),t}\|_2^4}\right]\right)^{1/4} \\ &\leq \frac{\|\epsilon_t\|_4}{\chi\|\epsilon_t\|_2}\left\|\mathcal{H}_{(i)}^{-1}\right\|_\infty^{1/4} \times \mathcal{M}_n^{1/4} \equiv \mathcal{K}\times\mathcal{M}_n^{1/4},\end{aligned}$$

where $\mathcal{K}$ is implicit. Hence, $u_{\epsilon s}(\gamma) = \mathcal{K}\mathcal{M}_n^{1/4}/\gamma$ in view of Markov and Liapunov inequalities:

$$\begin{aligned}P\left(\max_{1\leq i\leq k_{\theta,n},1\leq t\leq n}\left|\zeta_{(i),t}\right| > u_{\epsilon s}\right) &\leq \frac{1}{u_{\epsilon s}}E\left[\max_{1\leq i\leq k_{\theta,n},1\leq t\leq n}\left|\zeta_{(i),t}\right|\right] \\ &\leq \frac{1}{u_{\epsilon s}}\left(E\left[\max_{1\leq i\leq k_{\theta,n},1\leq t\leq n}\zeta_{(i),t}^2\right]\right)^{1/2} \leq \frac{1}{u_{\epsilon s}}\mathcal{K}\mathcal{M}_n^{1/4}.\end{aligned}$$

Second, use a log-exp argument with Jensen's inequality, and Gaussianicity, to yield

$$\begin{aligned}E\left[\max_{1\leq i\leq k_{\theta,n},1\leq t\leq n}\left|z_{(i),t}\right|\right] &\leq \frac{1}{\lambda}\ln\left(E\left[\exp\left(\lambda\max_{1\leq i\leq k_{\theta,n},1\leq t\leq n}\left|z_{(i),t}\right|\right)\right]\right) \text{ for any } \lambda > 0: \\ &\leq \frac{1}{\lambda}\ln\left(nk_{\theta,n}E\left[\exp\left\{\lambda\left|z_{(i),t}\right|\right\}\right]\right) \\ &= \frac{1}{\lambda}\ln\left(nk_{\theta,n}\right) + \lambda = 2\sqrt{\ln\left(nk_{\theta,n}\right)}.\end{aligned}$$

The final equality uses the minimizer $\sqrt{\ln(nk_{\theta,n})}$ of $\lambda^{-1}\ln(nk_{\theta,n}) + \lambda$. Hence $u_z(\gamma) = (2/\gamma)\sqrt{\ln(nk_{\theta,n})}$. Combine the above to yield:

$$u(\gamma) = \frac{1}{\gamma}\left\{\mathcal{K}\mathcal{M}_n^{1/4} \vee 2\sqrt{\ln(nk_{\theta,n})}\right\}.$$

Now use cases (i)-(iv) for $\mathcal{M}_n$ derived in the proof of Lemma 2.2 to deduce bounds for $u(\gamma)$, for some universal $\mathcal{C} > 0$: (i) $\mathcal{M}_n \leq K$ hence $u(\gamma) \leq \mathcal{C}\frac{1}{\gamma}\sqrt{\ln(nk_{\theta,n})}$; (ii) $\mathcal{M}_n \leq K\sqrt{\ln(k_{\theta,n})}$ hence $u(\gamma) \leq \frac{1}{\gamma}\mathcal{C}\sqrt{\ln(nk_{\theta,n})}$; (iii) $\mathcal{M}_n \leq K\ln(k_{\theta,n})$ hence again $u(\gamma) = \frac{1}{\gamma}\mathcal{C}\sqrt{\ln(nk_{\theta,n})}$; and (iv) $\mathcal{M}_n \leq Kk_{\theta,n}^{4/p}$ yields $u(\gamma) \leq \frac{1}{\gamma}\{\mathcal{K}k_{\theta,n}^{1/p} \vee 2\sqrt{\ln(nk_{\theta,n})}\} \leq \mathcal{C}\frac{1}{\gamma}k_{\theta,n}^{1/p}\sqrt{\ln(n)}$.

Next, return to (A.1). Let $\{g(n)\}$ be a positive monotonic sequence, $g(n) \to \infty$. Then $\rho_n(\lambda) = O(1/g(n)) = o(1)$ for any $\gamma = O(1/g(n))$ whenever $n^{-1/8}(\ln(k_{\theta,n}n/\gamma))^{7/8} \to 0$ and $n^{-1/2}(\ln(k_{\theta,n}n/\gamma))^{3/2}u(\gamma) \to 0$. Using $\gamma = O(1/g(n))$, the first condition reduces to

$$\ln(k_{\theta,n}) + \ln(ng(n)) = o(n^{1/7}) \Longrightarrow \ln(k_{\theta,n}) = o(n^{1/7}) \text{ if } \ln(g(n)) = o(n^{1/7}). \tag{A.2}$$

The second condition reduces to $n^{-1/2}[\ln(nk_{\theta,n}) + \ln(g(n)]^{3/2}u(\gamma) = o(1)$. Hence, for any slowly varying $g(n) \to \infty$, by the above case-specific bounds for $u(\gamma)$:

$$\begin{aligned} &(i)\text{-}(iii)\ \ln(nk_{\theta,n}) = o\left(n^{1/4}/g(n)^{1/2}\right) \Longrightarrow \ln(k_{\theta,n}) = o\left(\frac{n^{1/4}}{g(n)^{1/2}}\right) \\ &(iv)\ \ln(n)^{3/2}\left\{1 + \frac{\ln(k_{\theta,n}g(n))}{\ln(n)}\right\}^{3/2} k_{\theta,n}^{1/p} = o\left(\frac{n^{1/2}}{g(n)\sqrt{\ln(n)}}\right) \Longrightarrow k_{\theta,n} = o\left(\frac{n^{p/2}}{\{g(n)\ln^2(n)\}^p}\right) \end{aligned} \tag{A.3}$$

Now take the minimum $k_{\theta,n}$ from (A.2) and (A.3) to yield $\rho_n(\lambda) = O(1/g(n)) \to 0$ for slowly varying $g(n) \to \infty$ provided by case: (i)-(iii) $\ln(k_{\theta,n}) = o(n^{1/7})$, and (iv) $k_{\theta,n} = o(n^{p/2}/[g(n)\ln^2(n)]^p)$. $\mathcal{QED}$.

The proof of Theorem 2.5, and multiple lemmas in SM, exploit a general high dimensional probability bound applied to $\widehat{\mathcal{H}}_{(i)} - \mathcal{H}_{(i)}$, $\sqrt{n}\widehat{\mathcal{G}}_{(i)}$ and $\sqrt{n}(\hat{\beta}_{(i)} - \beta^*_{(i)})$. Recall $\widehat{\mathcal{G}}_{(i)} \equiv -1/n\sum_{t=1}^n v_{(i),t}x_{(i),t}$, $\widehat{\mathcal{H}}_{(i)} \equiv 1/n\sum_{t=1}^n x_{(i),t}x'_{(i),t}$, $\mathcal{H}_{(i)} \equiv E[x_{(i),t}x'_{(i),t}]$, and $\mathcal{M}_n \equiv E|x_{(i),t}|^4_\infty$.

Lemma A.1. *Let Assumption 1 hold, and let $\{k_{\theta,n}\}$ be any monotonically increasing sequence of integers. Then (a) $||\sqrt{n}\widehat{\mathcal{G}}_{(i)}||_\infty = O_p(\sqrt{\ln(k_{\theta,n})\mathcal{M}_n})$; (b) $||\widehat{\mathcal{H}}^{-1}_{(i)} - \mathcal{H}^{-1}_{(i)}||_\infty = O_p(\sqrt{\ln(k_{\theta,n})\mathcal{M}_n/n})$; and (c) $||\sqrt{n}(\hat{\beta}_{(i)} - \beta^*_{(i)})||_\infty = O_p(\sqrt{\ln(k_{\theta,n})\mathcal{M}_n})$ provided $\ln(k_{\theta,n}) = o(\sqrt{n}/\mathcal{M}_n)$.*

Remark A.4. $\ln(k_{\theta,n}) = o(\sqrt{n}/\mathcal{M}_n)$ holds under each covariate case $(i) - (iv)$: see the proof of Lemma 2.2 in SM. Hence $\ln(k_{\theta,n})\mathcal{M}_n/n \to 0$, yielding, e.g., $||\widehat{\mathcal{H}}^{-1}_{(i)} - \mathcal{H}^{-1}_{(i)}||_\infty \xrightarrow{p} 0$ and $||\hat{\beta}_{(i)} - \beta^*_{(i)}||_\infty \xrightarrow{p} 0$.

Proof of Theorem 2.5. For (a), convergence in distribution (7) follows from asymptotic expansion Lemma 2.2, and convergence in Kolmogorov distance (6), cf. Theorem 2.4 with $\lambda = [\mathbf{0}'_{k_\delta}, 1]'$. Then $|\sqrt{n}\mathcal{W}_{n,i}\hat{\theta}_i|_\infty \xrightarrow{d} \max_{i\in\mathbb{N}}|\mathcal{W}_i\boldsymbol{Z}_{(i)}([\mathbf{0}'_{k_\delta}, 1])|$ follows immediately. Claim (b) follows from Lemma A.1.c consistency of $\hat{\theta}_i$, and $\mathcal{W}_{n,i} \xrightarrow{p} \mathcal{W}_i$ with $\mathcal{W}_i \in (0, \infty)$. $\mathcal{QED}$.

Proof of Theorem 3.4.

Claim (a). We have $|\sqrt{n}\mathcal{W}_{n,i}\hat{\theta}_i|_\infty \overset{d}{\to} \max_{i\in\mathbb{N}}|\mathcal{W}_i\boldsymbol{Z}_{(i)}([\mathbf{0}'_{k_\delta},1])|$ from Theorem 2.5.a. Theorem 3.3 yields $\max_{1\leq i\leq k_{\theta,n}}\tilde{\mathcal{T}}_n \Rightarrow^p \max_{i\in\mathbb{N}}|\mathcal{W}_i\tilde{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])|$, where $\{\tilde{\boldsymbol{Z}}_{(i)}(\lambda)\}_{i\in\mathbb{N}}$ is an independent copy of $\{\boldsymbol{Z}_{(i)}(\lambda)\}_{i\in\mathbb{N}}$, $\boldsymbol{Z}_{(i)}(\lambda) \sim N(0, E[\epsilon_t^2]\lambda'\mathcal{H}_{(i)}^{-1}\lambda)$, independent of the asymptotic draw $\{x_t, y_t\}_{t=1}^\infty$. The proof now follows from arguments in Hansen (1996, p. 427).

Claim (b). Let H_1 hold, suppose $\theta_{0,j} \neq 0$ for some j, define $\mathfrak{S}_n \equiv \{x_t, y_t\}_{t=1}^n$, and write $\mathcal{P}_{\mathfrak{S}_n}(\mathcal{A}) \equiv \mathcal{P}(\mathcal{A}|\mathfrak{S}_n)$. In view of $\max_{1\leq i\leq k_{\theta,n}}\tilde{\mathcal{T}}_n \Rightarrow^p \max_{i\in\mathbb{N}}|\mathcal{W}_i\tilde{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])|$ we have (Gine and Zinn, 1990, eq. (3.4))

$$\sup_{z>0}\left|\mathcal{P}_{\mathfrak{S}_n}\left(\tilde{\mathcal{T}}_n \leq z\right) - P\left(\left|\mathcal{W}_i\mathring{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])\right|_\infty \leq z\right)\right| \overset{p}{\to} 0,$$

where $\{\mathring{\boldsymbol{Z}}_{(i)}(\cdot)\}_{i\in\mathbb{N}}$ is an independent copy of $\{\boldsymbol{Z}_{(i)}(\cdot)\}_{i\in\mathbb{N}}$. Moreover, $\tilde{p}_{n,\mathcal{R}_n} = \mathcal{P}_{\mathfrak{S}_n}(\tilde{\mathcal{T}}_{n,1} > \mathcal{T}_n) + o_p(1)$ by the Glivenko-Cantelli theorem and independence across bootstrap samples. Further, $\mathcal{T}_n \geq |\sqrt{n}\mathcal{W}_{n,j}(\hat{\theta}_j - \theta_{0,j}) + \sqrt{n}\mathcal{W}_{n,j}\theta_{0,j}| = \sqrt{n}\mathcal{W}_j\,|\theta_{0,j}|\,(|\hat{\theta}_j/\theta_{0,j}| + o_p(1)) \overset{p}{\to} \infty$ by Theorem 2.5.b and $|\mathcal{W}_{n,i} - \mathcal{W}_i|_\infty = o_p(1)$. Therefore

$$\tilde{p}_{n,\mathcal{R}_n} \leq P\left(\left|\mathcal{W}_i\mathring{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])\right|_\infty > \sqrt{n}\mathcal{W}_j\,|\theta_{0,j}|\left(\left|\frac{\hat{\theta}_j}{\theta_{0,j}}\right| + o_p(1)\right)\right) + o_p(1) \to 0$$

as long as $|\mathcal{W}_i\mathring{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])|_\infty = o_p(\sqrt{n})$. We show below that, given Gaussianicity of $\mathring{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])$:

$$\left|\mathcal{W}_i\mathring{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])\right|_\infty = O_p\left(\sqrt{\ln(k_{\theta,n})}\right) \tag{A.4}$$

hence $\tilde{p}_{n,\mathcal{R}_n} \overset{p}{\to} 0$ under H_1 and therefore $P(\tilde{p}_{n,\mathcal{R}_n} < \alpha) \to 1$ as claimed.

We now prove (A.4). Put $\zeta = \sqrt{2\mathcal{W}E[\epsilon_t^2]\mathcal{H}^{-1,\theta,\theta}} > 0$ where $\mathcal{W} \equiv \max_{i\in\mathbb{N}}\{\mathcal{W}_i\} < \infty$, and $\mathcal{H}^{-1,\theta,\theta} \equiv \max_{1\leq i\leq k_{\theta,n}}[\mathbf{0}'_{k_\delta},1]\mathcal{H}_{(i)}^{-1}[\mathbf{0}'_{k_\delta},1]' \in (0,\infty)$ under positive definiteness Assumption 1.c. Then Chernoff's Gaussian concentration bound $P(|Y| > c) \leq e^{-c^2/(2\sigma^2)}$ for $Y \sim N(0,\sigma^2)$, and Boole's inequality yield for any $\xi > 0$:

$$\begin{aligned} &P\left(\left|\mathcal{W}_i\mathring{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])\right|_\infty > \zeta\sqrt{\ln(k_{\theta,n})} + \xi\right) \\ &\quad\leq \sum_{i=1}^{k_{\theta,n}} P\left(\left|\mathcal{W}_i\mathring{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta},1])\right| > \zeta\sqrt{\ln(k_{\theta,n})} + \xi\right) \\ &\quad\leq \sum_{i=1}^{k_{\theta,n}} \exp\left\{-\frac{\left(\zeta\sqrt{\ln(k_{\theta,n})} + \xi\right)^2}{2\mathcal{W}E[\epsilon_t^2]\mathcal{H}^{-1,\theta,\theta}}\right\} \end{aligned} \tag{A.5}$$

$$= \frac{1}{k_{\theta,n}} \sum_{i=1}^{k_{\theta,n}} \exp\left\{ -\frac{2\xi\sqrt{\ln(k_{\theta,n})}}{\zeta} \right\} \exp\left\{ -\frac{\xi^2}{\zeta^2} \right\} \leq \exp\left\{ -\xi^2/\zeta^2 \right\}. \ \mathcal{QED}.$$

Proof of Theorem 3.5. Write $\xi_n \equiv \sqrt{n/\ln(k_{\theta,n})\,\mathcal{M}_n}$. Under H_1^L by the same arguments used to prove Lemma 2.2 and Theorem 2.5, and in view of Lemma A.1.c:

$$\left| \left| \sqrt{n}\mathcal{W}_{n,i}\left(\hat{\theta}_i - c_i\xi_n\right) \right|_\infty - \left| \mathcal{W}_i \boldsymbol{Z}_{(i)}([\mathbf{0}'_{k_\delta}, 1]) \right|_\infty \right| \overset{p}{\to} 0, \tag{A.6}$$

where $\boldsymbol{Z}_{(i)}([\mathbf{0}'_{k_\delta}, 1]) \sim N(0, E[\epsilon_t^2]\mathcal{H}^{-1,\theta,\theta})$ with $\mathcal{H}^{-1,\theta,\theta} \equiv \max_{1\leq i\leq k_{\theta,n}} [\mathbf{0}'_{k_\delta}, 1]\mathcal{H}_{(i)}^{-1}[\mathbf{0}'_{k_\delta}, 1]'$.

Now define for arbitrary $d = [d_i]_{i=1}^{k_{\theta,n}}$ and sample $\mathfrak{S}_n = \{x_t, y_t\}_{t=1}^n$:

$$\mathcal{P}_n(d) \equiv P\left(\frac{\left| \mathcal{W}_i \mathring{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta}, 1]) \right|_\infty}{\sqrt{\ln(k_{\theta,n})\,\mathcal{M}_n}} > \left| \xi_n \mathcal{W}_{n,i}\left(\hat{\theta}_i - c_i/\xi_n\right) + d_i \right|_\infty |\mathfrak{S}_n \right).$$

where $\{\mathring{\boldsymbol{Z}}_{(i)}(\cdot)\}_{i\in\mathbb{N}}$ is an independent copy of $\{\boldsymbol{Z}_{(i)}(\cdot)\}_{i\in\mathbb{N}}$. By the Glivenko-Cantelli Theorem with $\mathcal{R} = \mathcal{R}_n \to \infty$, the triangle inequality, $|\mathcal{W}_{n,i} - \mathcal{W}_i|_\infty = O_p(\sqrt{\ln(k_{\theta,n})\,\mathcal{M}_n/n})$, and the stated bounds on $k_{\theta,n}$ which ensure $\ln(k_{\theta,n}) = o(\sqrt{n}/\mathcal{M}_n)$ in all cases, and arguments in the proof of Theorem 3.4.b, the bootstrapped p-value approximation satisfies:

$$\begin{aligned}
\tilde{p}_{n,\mathcal{R}} &= P\left(\left| \sqrt{n}\mathcal{W}_{n,i}\widehat{\hat{\theta}}_i \right|_\infty > \left| \sqrt{n}\mathcal{W}_{n,i}\left(\hat{\theta}_i - c_i\xi_n\right) + \mathcal{W}_i c_i \sqrt{\ln(k_{\theta,n})\,\mathcal{M}_n} \right|_\infty |\mathfrak{S}_n \right) + o_p(1) \\
&= P\left(\frac{\left| \mathcal{W}_i \mathring{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta}, 1]) \right|_\infty}{\sqrt{\ln(k_{\theta,n})\,\mathcal{M}_n}} > \left| \xi_n \mathcal{W}_{n,i}\left(\hat{\theta}_i - c_i/\xi_n\right) + \mathcal{W}_i c_i \right|_\infty |\mathfrak{S}_n \right) + o_p(1) \\
&= \mathcal{P}_n\left(\mathcal{W}_j c_j\right) + o_p(1).
\end{aligned}$$

The proof of Theorem 3.4.a and arguments in Hansen (1996, p. 427) imply $\mathcal{P}_n(\mathbf{0}_{k_{\theta,n}})$ is asymptotically uniformly distributed.

Now, in view of (A.4) and (A.6) it follows that $\max_{1\leq i\leq k_{\theta,n}} |\mathcal{W}_i \mathring{\boldsymbol{Z}}_{(i)}([\mathbf{0}'_{k_\delta}, 1])| = O_p(\sqrt{\ln(k_{\theta,n})})$ and $\max_{1\leq i\leq k_{\theta,n}} |\xi_n \mathcal{W}_{n,i}(\hat{\theta}_i - c_i/\xi_n)| = O_p(1)$. Thus $\lim_{n\to\infty} P(\tilde{p}_{n,\mathcal{R}_n} < \alpha) \nearrow 1$ monotonically as the maximum local drift $\max_{i\in\mathbb{N}} |c_i| \nearrow \infty$ since $\mathcal{W}_i \in (0, \infty)\ \forall i$. $\mathcal{QED}$

References

Amemiya, T. (1980): “Selection of Regressors,” *International Economic Review*, 21, 331–354.

Andrews, D. W. K. (1991): “Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models,” *Econometrica*, 59, 307–345.

Arias-Castro, E., E. J. Candès, and Y. Plan (2011): “Global Testing Under Sparse

Alternative: ANOVA, Mulitple Comparisons and the Higher Criticism," *Annals of Statistics*, 39, 2533–2556.

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85, 233–298.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28, 29–50.

Bickel, P. J., and D. A. Freedman (1983): "Bootstrapping Regression Models with Many Parameters," in *A Festschrift for Erich L. Lehmann*, ed. by P. Bickel, K. Doksum, J. Hodges, and R. BocaRaton. Chapman and Hall.

Bose, A. (1988): "Edgewordth Correction by Bootstrap in Autoregressions," *Annals of Statistics*, 16, 1709–1722.

Bühlmann, P. (2013): "Statistical Significance in High-Dimensional Linear Models," *Bernoulli*, 19, 1212–1242.

Bühlmann, P., and S. Van De Geer (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg.

Cai, T., Z. Guo, and Y. Xia (2023): "Statistical Inference and Large-scale Multiple Testing for High-Dimensional Regression Models," *Test*, forthcoming.

Cai, T. T., and Z. Guo (2017): "Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity," *Annals of Statistics*.

Carpentier, A., and N. Verzelen (2019): "Adaptive Estimation of the Sparsity in the Gaussian Vector Model," *Annals of Statistics*, 47, 93–126.

Castle, J., X. Qin, and W. R. Reed (2009): "How To Pick The Best Regression Equation: A Review And Comparison Of Model Selection Algorithms," Discussion Paper 13/2009, Dept. of Economics and Finance, University of Canterbury.

Cattaneo, M. D., M. Jansson, and W. Newey (2018): "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity," *Journal of the American Statistical Association*, 113, 1350–1361.

Chen, X., and D. Pouzo (2015): "Sieve Wald and QLR Inference on Semi-Nonparametric Conditional Moment Models," *Econometrica*, 83, 1013–1079.

Chen, X., M. Xu, and W. B. Wu (2013): "Covariance and Precision Matrix Estimation for High-Dimensional Time Series," *Annals of Statistics*, 41, 2994–3021.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2016): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, 21, C1–C68.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013): "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors," *Annals of Statistics*, 41, 2786–2819.

——— (2017): "Central Limit Theorems and Bootstrap in High Dimensions," *Annals of Probability*, 45, 2309–2352.

Davidson, R., and J. G. MacKinnon (2006): "The Power of Bootstrap and Asymptotic Tests," *Journal of Econometrics*, 133, 421–441.

Dezeure, R., P. Bühlmann, and C.-H. Zhang (2017): "High-Dimensional Simultaneous Inference with the Bootstrap," *Test*, 26, 685–719.

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

Fan, J., J. Lv, and . L. Qi (2011): "Sparse High-Dimensional Models in Economics," *Annual Economic Review*, 3, 291–317.

Fan, J., and H. Peng (2004): "Nonconcave Penalized Likelihood with a Diverging Number of Parameters," *Annals of Statistics*, 32, 928–961.

Gautier, E., and C. Rose (2021): "High-Dimensional Instrumental Variables Regression and Confidence Sets," hal-00591732v7.

Genovese, C. R., J. Jin, L. Wasserman, and Z. Yao (3023): "A Comparison of the Lasso and Marginal Regression," *Journal of Machine Learnging Research*, 13, 2107–2143.

Ghysels, E., J. B. Hill, and K. Motegi (2020): "Testing a Large Set of Zero Restrictions in Regression Models, with an Application to Mixed Frequency Granger Causality," *Journal of Econometrics*, 218, 633–654.

Gine, E., and J. Zinn (1990): "Bootstrapping General Empirical Measures," *Annals of Probability*, 18, 851–869.

Gonçalves, S., and L. Kilian (2004): "Bootstrapping Autoregressions with Conditional Heteroskedasticity of Unknown Form," *Journal of Econometrics*, 123, 89–120.

Hansen, B. E. (1996): "Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis," *Econometrica*, 64, 413–430.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics. Springer, New York, 2nd edn.

Hill, J. B. (2023): "Supplemental Material for *Testing Many Zero Restrictions in a High Dimensional Linear Regression Setting*," Dept. of Economics, University of North Carolina.

Huang, J., J. L. Horowitz, and S. Ma (2008): "Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models," *Annals of Statistics*, 36, 587–613.

Kabe, D. G. (1963): "Stepwise Multivariate Linear Regression," *Journal of the American Statistical Association*, 58, 770–773.

Li, C., and U. K. Müller (2021): "Linear Regression with Many Controls of Limited Explanatory Power," *Quantitative Economics*, 12, 405–442.

Liu, R. Y. (1988): "Bootstrap Procedures under some Non-I.I.D. Models," *Annals of Statistics*, 16, 1696–1708.

Liu, W., and S. Luo (2014): "Hypothesis Testing for High-Dimensional Regression Models," Technical Report.

Mammen, E. (1993): "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *Annals of Statistics*, 21, 255–285.

McKeague, I., and M. Qian (2015): "An Adaptive Resampling Test for Detecting the Presence of Significant Predictors," *Journal of the American Statistical Association*, 110, 1422–1433.

McKeague, I., and I. Zhang (2022): "Significance Testing for Canonical Correlation Analysis in High Dimensions," *Biometrika*, 109, 1076–1083.

Royston, P., and D. G. Altman (1994): "Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling," *Journal of the Royal Statistical Society, Series B*, 43, 429–467.

Sawa, T. (1978): "IInformation Criteria for Discriminating Among Alternative Regression Models," *Econometrica*, 46, 1273–1291.

Stone, C. J. (1981): "Admissible Selection of an Accurate and Parsimonious Normal Linear Regression Model," *Annals of Statistics*, 9, 475–485.

Vershynin, R. (2018): *High-Dimensional Probability*. Cambridge University Press, Cambridge, UK.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

Zhang, C.-H., and S. S. Zhang (2014): "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society, Series B*, 76, 217–242.

Zhang, D., and W. B. Wu (2017): "Gaussian Approximation for High Dimensional Time Series," *Annals of Statistics*, 45, 1895–1919.

Zhong, P.-S., S. X. Chen, and M. Xu (2013): "Tests Alternative to Higher Criticism High-Dimensional Mean under Sparsity and Column-Wise Dependence," *Annals of Statistics*, 41, 2820–2851.

Table 2: Rejection Frequencies under H_0 and $H_1(i)$

	$H_0 : \theta_0 = 0$									$H_1(i) : \theta_{0,1} = .0015$ with $[\theta_{0,i}]_{i=2}^{k_{\theta,n}} = \mathbf{0}_{k_{\theta,n}-1}$								
	$n = 100$									$n = 100$								
	$k_{\theta,n} = 35$			$k_{\theta,n} = 200$			$k_{\theta,n} = 482$			$k_{\theta,n} = 35$			$k_{\theta,n} = 200$			$k_{\theta,n} = 482$		
Test / Size	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
p-Max-Test	.010	.042	.116	.013	.053	.103	.008	.055	.096	.923	.981	.993	1.00	1.00	1.00	1.00	1.00	1.00
p-Max-t-Test	.007	.058	.111	.013	.061	.117	.012	.059	.112	.946	.986	.994	1.00	1.00	1.00	1.00	1.00	1.00
dbl-Max-Test	.014	.054	.108	.000	.022	.068	.021	.078	.176	.012	.048	.092	.020	.044	.100	.036	.180	.332
dbl-Max-t-Test	.014	.032	.082	.004	.044	.090	.028	.084	.196	.016	.056	.084	.132	.188	.220	.804	.900	.904
Wald	.000	.000	.018	-	-	-	-	-	-	.002	.113	.341	-	-	-	-	-	-
	$n = 250$									$n = 250$								
	$k_{\theta,n} = 35$			$k_{\theta,n} = 1144$			$k_{\theta,n} = 1250$			$k_{\theta,n} = 35$			$k_{\theta,n} = 1144$			$k_{\theta,n} = 1250$		
	1%	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
p-Max-Test	.010	.056	.106	.012	.058	.112	.007	.046	.095	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
p-Max-t-Test	.012	.055	.106	.015	.061	.116	.009	.057	.097	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Wald	.000	.015	.066	-	-	-	-	-	-	.961	.996	1.00	-	-	-	-	-	-
	$n = 500$									$n = 500$								
	$k_{\theta,n} = 35$			$k_{\theta,n} = 2381$			$k_{\theta,n} = 2500$			$k_{\theta,n} = 35$			$k_{\theta,n} = 2381$			$k_{\theta,n} = 2500$		
Test / Size	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
p-Max-Test	.012	053	.105	.011	.055	.103	.007	.052	.103	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
p-Max-t-Test	.012	.051	.107	.009	.054	.111	.009	.054	.111	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Wald	.003	.043	.076	-	-	-	-	-	-	.998	1.00	1.00	-	-	-	-	-	-

p = parsimonious; *dbl* = de-biased Lasso. x_t are unbounded. $k_\delta = 0$. All test p-values are bootstrapped, based on 1,000 (p-max, Wald) or 250 (*dbl*-max) independently drawn samples.

Table 3: Rejection Frequencies under $H_1(ii)$ and $H_1(iii)$

	$H_1(ii): \theta_{0,i} = .0002i/k_{\theta,n}$ for $i = 1, ..., k_{\theta,n}$									$H_1(iii): \theta_{0,i} = .00015$ for $i = 1, ..., 10$								
	$n = 100$									$n = 100$								
	$k_{\theta,n} = 35$			$k_{\theta,n} = 200$			$k_{\theta,n} = 482$			$k_{\theta,n} = 35$			$k_{\theta,n} = 200$			$k_{\theta,n} = 482$		
Test / Size	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
p-Max-Test	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.934	.988	.993	1.00	1.00	1.00	1.00	1.00	1.00
p-Max-t-Test	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.954	.989	.993	1.00	1.00	1.00	1.00	1.00	1.00
dbl-Max-Test	.016	.060	.116	.188	.436	.596	.468	.748	.864	.016	.048	.092	.000	.036	.084	.020	.072	.120
dbl-Max-t-Test	.008	.036	.088	.380	.536	.584	.503	.810	.900	.012	.036	.076	.000	.008	.016	.016	.024	.028
Wald	.542	.982	.998	-	-	-	-	-	-	.003	.101	.307	-	-	-	-	-	-
	$n = 250$									$n = 250$								
	$k_{\theta,n} = 35$			$k_{\theta,n} = 1144$			$k_{\theta,n} = 1250$			$k_{\theta,n} = 35$			$k_{\theta,n} = 1144$			$k_{\theta,n} = 1250$		
Test / Size	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
p-Max-Test	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.998	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
p-Max-t-Test	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Wald	1.00	1.00	1.00	-	-	-	-	-	-	.730	.944	.980	-	-	-	-	-	-
	$n = 500$									$n = 500$								
	$k_{\theta,n} = 35$			$k_{\theta,n} = 2381$			$k_{\theta,n} = 2500$			$k_{\theta,n} = 35$			$k_{\theta,n} = 2381$			$k_{\theta,n} = 2500$		
Test / Size	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
p-Max-Test	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
p-Max-t-Test	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Wald	1.00	1.00	1.00	-	-	-	-	-	-	1.00	1.00	1.00	-	-	-	-	-	-

p = parsimonious; *dbl* = de-biased Lasso. x_t are unbounded. $k_\delta = 0$. All test p-values are bootstrapped, based on 1,000 (p-max, Wald) or 250 (*dbl*-max) independently drawn samples.