Learning Before Testing: A Selective Nonparametric Test for Conditional Moment Restrictions[*]

Jia Li[†] Zhipeng Liao[‡] Wenyu Zhou[§]

January 20, 2022

Abstract

This paper develops a new test for conditional moment restrictions via nonparametric series regression, with approximating series terms selected by Lasso. Machine-learning the main features of the unknown conditional expectation function beforehand enables the test to seek power in a targeted fashion. The data-driven selection, however, also tends to distort the test's size nontrivially, because it restricts the (growing-dimensional) score vector in the series regression on a random polytope, and hence, effectively alters the score's asymptotic normality. A novel critical value is proposed to account for this truncation effect. We establish the size and local power properties of the proposed selective test under a general setting for heterogeneous serially dependent data. The local power analysis reveals a desirable adaptive feature of the test in the sense that it may detect smaller deviations from the null when the unknown function is less complex. Monte Carlo evidence demonstrates the superior finite-sample size and power properties of the proposed test relative to some benchmarks.

Keywords: conditional moments; Lasso; machine learning; series estimation; uniform inference; variable selection.
JEL Codes: C14, C22.

[*] We thank Francis Diebold, Xu Cheng, Bruce Hansen, Jack Porter, Frank Schorfheide, Xiaoxia Shi, Kenneth West, Zhenlin Yang, Jun Yu and seminar participants at the Chinese University of Hong Kong, Singapore Management University, University of Wisconsin-Madison, University of Pennsylvania for their comments.

[†] School of Economics, Singapore Management University, Singapore; e-mail: jiali@smu.edu.sg.

[‡] Department of Economics, UCLA, Log Angeles, CA 90095; e-mail: zhipeng.liao@econ.ucla.edu.

[§] International Business School, Zhejiang University, Haining, Zhejiang 314400, China; e-mail: wenyuzhou@intl.zju.edu.cn.

1 Introduction

Testing conditional moment restrictions is an important topic in econometrics. One approach is to nonparametrically estimate the conditional moment function via a series regression (Andrews (1991a), Newey (1997)) and then test whether the function is zero in a uniform sense (Belloni, Chernozhukov, Chetverikov, and Kato (2015), Li and Liao (2020)). In practice, however, it is often difficult to decide which series terms should be employed to approximate the unknown function: Using too few may induce bias, but using too many hurts efficiency. To address this issue, it seems natural to apply some machine-learning-based variable selection procedure such as the Lasso (Tibshirani (1996)) and its variants. Although such methods may achieve the so-called "oracle property" in large samples, they are unlikely to completely meet that theoretical ideal in finite samples. Ignoring the sampling variability in the selection step may thus lead to possibly severe size distortions in the subsequent test (see Section 4 for concrete Monte Carlo evidence).

The main contribution of this paper is to propose a new critical value for the nonparametric test, which properly accounts for the effect of the preliminary Lasso-based selection. Our analysis reveals that the first-stage selection affects the second-stage inference by restricting the series-regression score on a random polytope. Since the asymptotics of the series estimator is captured by the (growing-dimensional) Gaussian coupling for the score vector, this restriction effectively results in a form of truncated normality, which explains the size distortion of the "naive" critical value directly constructed from the conventional asymptotic Gaussian approximation. The novel critical value proposed in this paper accounts for the truncation effect, and it greatly improves the test's size control in finite samples as shown in our simulation study.

To better understand the power property of the proposed test, we also characterize local alternatives against which the test is consistent. The power analysis clarifies an adaptive feature of the proposed selective test: The test is able to detect smaller deviations from the null if the deviation has a simpler form. In the extreme case when the unknown function can be approximated by a bounded number of series terms (but with a priori unknown identities), the test achieves consistency nearly—up to a logarithmic factor—at the parametric rate. In the worst-case scenario in which the unknown function is "very complex" (e.g., all covariates have equal predictive power), the power of the selective test deteriorates to the same level as the benchmark non-selective test. But in general, the former is shown to be more powerful than the latter.

The remainder of this paper is organized as follows. We present the econometric method and the related asymptotic theory in Section 2, followed by a couple of extensions in Section 3. Section 4 demonstrates the finite-sample performance of the proposed test in a Monte Carlo experiment. Section 5 concludes. The Appendix provides requisite implementation details and a pedagogical example. Technical proofs and additional simulation results are relegated to the Online Supplemental Appendix.

2 Main theory

We present our main theory in this section. Section 2.1 describes the testing problem and some related background. Section 2.2 presents the new selective test, with its theoretical properties established in Section 2.3.

2.1 The testing problem and some background

We start with introducing the econometric setting. Consider a series $(Y_t, X_t^\top)$, $1 \leq t \leq n$, of observed data, where Y_t is scalar-valued and X_t takes values in a compact set $\mathcal{X} \subseteq \mathbb{R}^d$.[1] Denote the conditional expectation function of Y_t given X_t by

$$g(x) \equiv \mathbb{E}\left[Y_t | X_t = x\right], \quad x \in \mathcal{X},$$

with the associated residual term $\epsilon_t \equiv Y_t - g\left(X_t\right)$. Our econometric interest is to test the null hypothesis

$$H_0 : g(x) = 0 \text{ for all } x \in \mathcal{X}, \tag{2.1}$$

against its complementary alternative, that is, $g(x) \neq 0$ for some x. In some applications, Y_t may depend on an unknown finite-dimensional parameter θ^* and, if so, we may emphasize this dependence explicitly by writing $Y_t\left(\theta^*\right)$. The testing of conditional moment restrictions arises routinely from various empirical settings. To help fix ideas, we briefly consider two prototype examples.

Example 1 (Forecast Rationality). Suppose that at time t a forecaster produces a one-period-ahead forecast $F_{t+1|t}$ for the next period's target variable, denoted $F_{t+1}^\dagger$ (e.g., inflation). It is well-known that given a time-t information set $\mathcal{I}_t$, the optimal forecast that minimizes the mean-squared-error loss is the conditional expectation $\mathbb{E}[F_{t+1}^\dagger | \mathcal{I}_t]$. Therefore, if $F_{t+1|t}$ is optimal, the forecasting error $Y_t = F_{t+1|t} - F_{t+1}^\dagger$ should satisfy $\mathbb{E}\left[Y_t | X_t\right] = 0$ for any X_t in the $\mathcal{I}_t$ information

[1] We consider scalar-valued Y_t mainly for ease of exposition. The econometric method can be trivially extended to accommodate multivariate Y_t.

set. In this setting, a test for (2.1) can be interpreted as a test for forecast rationality, as studied by Hansen and Hodrick (1980), Brown and Maital (1981), and Romer and Romer (2000), among others.

Example 2 (Euler and Bellman Equations). In dynamic equilibrium models, the equilibrium is often characterized by Euler equations in the form of conditional moment restrictions. The classical example is Hansen and Singleton's (1982) study of consumption-based asset pricing, in which the one-period-ahead pricing equation takes the form

$$\mathbb{E}\left[\left.\frac{\beta U'\left(C_{t+1},\gamma\right)}{U'\left(C_{t},\gamma\right)}R_{t+1}-1\right|X_t\right]=0, \tag{2.2}$$

where $U'\left(\cdot,\gamma\right)$ is the representative agent's marginal utility function with a preference parameter γ, β is the discounting factor, C_t is the consumption process, R_{t+1} is the return of an asset, and X_t is the state variable underlying the dynamic model. Equation (2.2) can be written in the form of (2.1) by setting $\theta^* = (\beta,\gamma)$ and $Y_t\left(\theta^*\right) = \frac{\beta U'(C_{t+1},\gamma)}{U'(C_t,\gamma)}R_{t+1}-1$. Similar equilibrium conditions can also be derived from Bellman equations; see, for example, Li and Liao (2020). In the macroeconomic setting, the parameter θ^* is often, though not always, calibrated based on external data and auxiliary models.

As seen from these examples, the Y_t variable may play different roles in different contexts and sometimes may involve a finite-dimensional parameter θ^* that may be estimated or calibrated depending on the style of empirical research. In addition, it is generally important to accommodate time-series dependence for these applications. In the remainder of Section 2, we shall assume that Y_t is directly observed so as to focus on the main innovation of the present paper. It is relatively straightforward to allow for the presence of an unknown θ^*, and we will develop that extension in Section 3.

Testing the hypothesis in (2.1) is econometrically nontrivial, as it concerns the global property of the conditional expectation function $g\left(\cdot\right)$. In practice, empiricists often take "parametric shortcuts" to bypass that nonparametric functional inference problem. The simplest way to do so is to integrate out the conditioning variable X_t and test the unconditional moment restriction $\mathbb{E}\left[Y_t\right]=0$. This amounts to regressing Y_t on a constant term and then conducting a t-test. To incorporate the conditioning information in X_t, it is common to run a linear regression

$$Y_t = a + b^\top X_t + e_t, \tag{2.3}$$

and then test whether the coefficients are all zero. The pros and cons of the parametric approach are also well understood. On one hand, if the observed data were known to be generated under the conjectured specification, the parametric approach would clearly be the simplest and the most efficient way to carry out the test. On the other hand, if the null is violated in a way that is "orthogonal" to the given parametric specification, the test will have little power in detecting it. In applied work, non-rejections may thus be challenged by a critical reader, because the parametric test is designed to seek power only in very specific directions that are generally hard to justify on the ground of economic theory.

A natural way to address this issue is to make the regression more flexible by including additional nonlinear terms. Following Andrews (1991a) and Newey (1997), one may formalize this more general approach as a nonparametric series regression so as to directly attack the functional inference. Consider a collection of approximating basis functions $(p_j(\cdot))_{1\leq j\leq m}$ such as polynomials, splines, trigonometric functions, wavelets, etc., and set $P(\cdot) \equiv (p_1(\cdot), \ldots, p_m(\cdot))^\top$. We may regress Y_t on $P(X_t)$ and construct the associated nonparametric series estimator for $g(\cdot)$ as

$$\hat{g}(\cdot) \equiv P(\cdot)^\top \left(\sum_{t=1}^{n} P(X_t) P(X_t)^\top \right)^{-1} \left(\sum_{t=1}^{n} P(X_t) Y_t \right). \tag{2.4}$$

With the number of series terms $m \to \infty$, the specification of this regression becomes increasingly more flexible in larger samples, so that the series approximation will approach the true unknown function. The test can then be carried out by examining whether the estimated function $\hat{g}(\cdot)$ is statistically zero in a uniform sense. A theoretical subtlety stems from the fact that the uniform inference for the series estimator is a non-Donsker problem due to the growing dimensionality of the regressors.[2] Li and Liao (2020) show that the estimation error function $\hat{g}(\cdot) - g(\cdot)$ can be strongly approximated, or coupled, by a sequence of divergent Gaussian processes in a time-series setting for general heterogeneous mixingales; also see Belloni, Chernozhukov, Chetverikov, and Kato (2015) for a similar analysis in the cross-sectional setting. Based on that theory, one may test $g(\cdot) = 0$ using a "functional t-test" based on the sup-t statistic $\sup_{x\in\mathcal{X}} |\hat{g}(x)| / \hat{s}(x)$, where $\hat{s}(\cdot)$ is an estimator of the standard error function of $\hat{g}(\cdot)$. The null hypothesis is rejected if the sup-t statistic is greater than a critical value determined by the strong Gaussian approximation. A more detailed discussion is given in Section 2.2 below.

The advantage of the nonparametric approach is that it speaks directly to the original hypothesis (2.1), whereas the parametric approach concentrates only on some of its implications. That being said, the flexibility of the nonparametric approach comes with an efficiency cost, which manifests theoretically in the relatively slow rate of convergence of the nonparametric estimator. The cost is also easily understood in practical terms. Indeed, if $g(x)$ happens to be a linear function in

[2] That is, the $\hat{g}(\cdot)$ estimator does not satisfy a functional central limit theorem (i.e., Donsker theorem) as considered in Pollard (2001), Andrews (1994), and van der Vaart and Wellner (1996), among others.

x, adding (a growing number of) higher-order polynomial terms in the series regression ought to be a "waste." In practice, the null hypothesis may be rejected by a simple parametric test, whereas the theoretically "omniscient" nonparametric test may be less powerful and fail to reject.

The above discussion clarifies the trade-off between flexibility and efficiency in the present testing context: Flexibility requires the inclusion of a large number of regressors that might be useful, but in order to achieve sharper inference, it would be better to focus on a small number of regressors that are actually useful for capturing the main features of the alternative. Our goal is to improve this trade-off margin. A reasonable approach is to first properly select a subset of approximating functions in the spirit of "feature extraction," and only use them to run the series regression and construct the sup-t test statistic. Given this goal, as well as the least-squares structure of the series regression, the Lasso method is clearly the most natural choice for implementing the selection.

We refer to this proposal as the *selective test*. The aforementioned construction of the test statistic is arguably straightforward in view of the prior literature on series estimation and Lasso. The key challenge for carrying out the selective test, however, is to properly determine its critical value. This turns out to be highly nonstandard. A seemingly reasonable approach is to treat the Lasso-selected subset of approximating functions "as given" (i.e., ignoring the fact that it is data-driven), and compute the critical value in the same way as in the benchmark non-selective test (Belloni, Chernozhukov, Chetverikov, and Kato (2015), Li and Liao (2020)). However, as we shall show in the simulations (see Section 4), this approach may result in quite nontrivial size distortions. This motivates us to develop a correction for the critical value so as to account for the selection effect. Our analysis is in spirit akin to the sequential inference theory commonly seen in econometrics (cf. Section 6 of Newey and McFadden (1994)). But, unlike the conventional setting, our "first-stage estimator" may be viewed as *set-valued* (in the form of a selection event), and it affects the second-stage through a fairly complicated truncation of the support of the coupling Gaussian process that drives the asymptotics of the series estimator. We now turn to the details.

2.2 The selective test

We construct the selective test in this subsection. Let $\mathcal{M}$ denote the index set associated with a collection $(p_j(\cdot))_{1\leq j\leq m}$ of candidate approximating functions. For ease of discussion, we identify $\mathcal{M}$ with the associated collection of approximating functions, and refer to it as a *dictionary*. We assume that the size of $\mathcal{M}$ grows asymptotically (i.e., $m \to \infty$ as $n \to \infty$), though its dependence on n is kept implicit in our notation for simplicity. For any nonempty subset $\mathcal{S} \subseteq \mathcal{M}$, we denote $P_{\mathcal{S}}(\cdot) \equiv (p_j(\cdot))_{j\in\mathcal{S}}$, which collects a subset of approximating functions selected by $\mathcal{S}$. The specific ordering of the $p_j(\cdot)$ components is irrelevant, because our statistics of interest are all invariant to the ordering. We refer to $\mathcal{S}$ as a *selection* and denote its cardinality by $|\mathcal{S}|$. Analogous to (2.4), the series estimator for $g(\cdot)$ based on the selection $\mathcal{S}$ is given by

$$\hat{g}_{\mathcal{S}}(\cdot) \equiv P_{\mathcal{S}}(\cdot)^\top \left(\sum_{t=1}^{n} P_{\mathcal{S}}(X_t) P_{\mathcal{S}}(X_t)^\top\right)^{-1} \left(\sum_{t=1}^{n} P_{\mathcal{S}}(X_t) Y_t\right). \tag{2.5}$$

Note that this includes the non-selective estimator $\hat{g}(\cdot)$ defined in (2.4) as a special case corresponding to $\mathcal{S} = \mathcal{M}$. The standard error function associated with $\hat{g}_{\mathcal{S}}(\cdot)$ is given by

$$\sigma_{\mathcal{S}}(\cdot) \equiv \sqrt{P_{\mathcal{S}}(\cdot)^\top Q_{\mathcal{S}}^{-1} \Sigma_{\mathcal{S}} Q_{\mathcal{S}}^{-1} P_{\mathcal{S}}(\cdot)},$$

where $Q_{\mathcal{S}} \equiv n^{-1}\sum_{t=1}^{n} \mathbb{E}\left[P_{\mathcal{S}}(X_t)P_{\mathcal{S}}(X_t)^\top\right]$ and $\Sigma_{\mathcal{S}} \equiv \mathrm{Var}[n^{-1/2}\sum_{t=1}^{n} P_{\mathcal{S}}(X_t)\epsilon_t]$. We may estimate $Q_{\mathcal{S}}$ via $\widehat{Q}_{\mathcal{S}} \equiv n^{-1}\sum_{t=1}^{n} P_{\mathcal{S}}(X_t)P_{\mathcal{S}}(X_t)^\top$ and estimate $\Sigma_{\mathcal{S}}$ using a (growing-dimensional) heteroskedasticity and autocorrelation consistent (HAC) estimator $\widehat{\Sigma}_{\mathcal{S}}$, following known results in the literature.[3] The standard error function $\sigma_{\mathcal{S}}(\cdot)$ can then be estimated via

$$\widehat{\sigma}_{\mathcal{S}}(\cdot) \equiv \sqrt{P_{\mathcal{S}}(\cdot)^\top \widehat{Q}_{\mathcal{S}}^{-1} \widehat{\Sigma}_{\mathcal{S}} \widehat{Q}_{\mathcal{S}}^{-1} P_{\mathcal{S}}(\cdot)}. \tag{2.6}$$

Finally, we define the sup-t test statistic associated with the selection $\mathcal{S}$ as

$$\widehat{T}_{\mathcal{S}} \equiv \sup_{x\in\mathcal{X}} \left|\frac{n^{1/2}\hat{g}_{\mathcal{S}}(x)}{\widehat{\sigma}_{\mathcal{S}}(x)}\right|. \tag{2.7}$$

We use Lasso to perform a data-driven selection from the dictionary $\mathcal{M}$. In some empirical applications, the user may consider a subset $\mathcal{M}_0 \subseteq \mathcal{M}$ of regressors to be important a priori (possibly based on economic reasoning) and like to "manually" select them into the series regression. To accommodate this type of customization, we design a selection procedure that always includes the *prior choice set* $\mathcal{M}_0$ and relies on Lasso to select additional regressors from the remainder set $\mathcal{M}_0^c \equiv \mathcal{M} \setminus \mathcal{M}_0$. For example, the user may insist on using a constant and a linear term in the series regression, but is uncertain about which higher-order polynomial terms should be included in addition. In this situation, they may put the constant and linear terms in $\mathcal{M}_0$, and let Lasso to "machine-learn" whether and which additional terms are needed. The role of Lasso in this design is thus to assist the user's choice rather than dictating it. Below, we maintain a mild convention

[3] We may take $\widehat{\Sigma}_{\mathcal{S}}$ to be the classical Newey–West estimator or more generally the HAC estimators studied by Andrews (1991b). The consistency and rate of convergence of these HAC estimators have been established in a general time-series setting with growing dimensions by Li and Liao (2020); see their Lemma B3. The consistency and rate of convergence of $\widehat{Q}_{\mathcal{S}}$ towards $Q_{\mathcal{S}}$ follow a law of large numbers of growing-dimensional matrices; see, for example, Lemma 2.2 in Chen and Christensen (2015) and Lemma B2 in Li and Liao (2020).

that $\mathcal{M}_0$ contains at least the constant term (which is also our recommended default choice); this ensures the selected set of regressors to be non-empty, and hence, avoids an uninteresting degeneracy.

This "Lasso-assisted" selection is implemented as follows. Given the user's prior choice $\mathcal{M}_0$, we perform a Lasso estimation with the resulting estimator given by

$$\hat{\beta}^{Lasso} \equiv \underset{\beta\in\mathbb{R}^m}{\operatorname{argmin}} \left\{ \frac{1}{2}\sum_{t=1}^{n}(Y_t - P(X_t)^\top \beta)^2 + \lambda_n \sum_{j\in\mathcal{M}_0^c} \omega_j \left|\beta_j\right| \right\}, \tag{2.8}$$

where λ_n is a sequence of penalty parameters commonly seen in Lasso-type problems, and $(\omega_j)_{j\in\mathcal{M}_0^c}$ is a collection of positive weights. Note that the L_1 penalty is applied only to the remainder set $\mathcal{M}_0^c$, whereas the coefficients in the prior choice set $\mathcal{M}_0$ are unrestricted. A simple choice of the ω_j weights is to set $\omega_j = 1$ identically or $\omega_j = \sqrt{n^{-1}\sum_{t=1}^{n} p_j(X_t)^2}$ (see, e.g., Zhao and Yu (2006), Bickel, Ritov, and Tsybakov (2009), and Belloni and Chernozhukov (2011)), but the more general setting in (2.8) also accommodates the adaptive Lasso (Zou (2006)). In Appendix A.1, we provide a concrete data-driven choice of these penalty parameters and establish its theoretical validity within our econometric framework. Due to the L_1 penalty, many coefficients indexed by $\mathcal{M}_0^c$ will be shrunk to zero. Our Lasso-assisted selection is then given by

$$\mathcal{L} \equiv \mathcal{M}_0 \bigcup \left\{ j \in \mathcal{M}_0^c : \hat{\beta}_j^{Lasso} \neq 0 \right\}, \tag{2.9}$$

which consists of the user's ex ante choice $\mathcal{M}_0$ and Lasso's ex post selection from $\mathcal{M}_0^c$. The corresponding selective test statistic is defined as

$$\widehat{T}_{\mathcal{L}} \equiv \widehat{T}_{\mathcal{S}}\big|_{\mathcal{S}=\mathcal{L}} = \sup_{x\in\mathcal{X}} \left| \frac{n^{1/2}\hat{g}_{\mathcal{L}}(x)}{\widehat{\sigma}_{\mathcal{L}}(x)} \right|. \tag{2.10}$$

A large value of the test statistic $\widehat{T}_{\mathcal{L}}$ signifies a violation of the null hypothesis (i.e., $g(\cdot) \neq 0$).

The remaining task is to determine a critical value for $\widehat{T}_{\mathcal{L}}$. Before elaborating our proposal, it is instructive to first review how the critical value may be constructed in a simpler benchmark scenario with a nonrandom selection $\mathcal{S}$, as studied in Belloni, Chernozhukov, Chetverikov, and Kato (2015) and Li and Liao (2020).[4] These authors show that $\widehat{T}_{\mathcal{S}}$ can be strongly approximated by the supremum of a Gaussian process under the null hypothesis. More precisely, there exists a sequence of $|\mathcal{S}|$-dimensional Gaussian random vectors $\widetilde{N}_{\mathcal{S}} \sim \mathcal{N}(0, \Sigma_{\mathcal{S}})$ such that

$$\widehat{T}_{\mathcal{S}} - \widetilde{T}_{\mathcal{S}} = o_p(1), \quad \text{where} \quad \widetilde{T}_{\mathcal{S}} \equiv \sup_{x\in\mathcal{X}} \left| \frac{P_{\mathcal{S}}(x)^\top Q_{\mathcal{S}}^{-1}\widetilde{N}_{\mathcal{S}}}{\sigma_{\mathcal{S}}(x)} \right|. \tag{2.11}$$

[4] The theory of Belloni, Chernozhukov, Chetverikov, and Kato (2015) and Li and Liao (2020) does not explicitly involve selection, and hence, corresponds to the case with $\mathcal{S} = \mathcal{M}$. But it is easy to see that their inference theory can be trivially adapted to accommodate any nonrandom selection $\mathcal{S}$.

The $1-\alpha$ quantile of $\widetilde{T}_{\mathcal{S}}$ can thus be used as a critical value for $\widehat{T}_{\mathcal{S}}$ at significance level α. A feasible version of this critical value can be estimated as the $1-\alpha$ quantile of

$$\widetilde{T}_{\mathcal{S}}^{*} \equiv \sup_{x \in \mathcal{X}} \left| \frac{P_{\mathcal{S}}(x)^{\top} \widehat{Q}_{\mathcal{S}}^{-1} \widetilde{N}_{\mathcal{S}}^{*}}{\widehat{\sigma}_{\mathcal{S}}(x)} \right|, \tag{2.12}$$

where $\widetilde{N}_{\mathcal{S}}^{*}$, conditional on data, is $\mathcal{N}(0, \widehat{\Sigma}_{\mathcal{S}})$-distributed.

A seemingly natural way to construct $\widehat{T}_{\mathcal{L}}$'s critical value is to directly apply this benchmark theory by plugging in $\mathcal{S} = \mathcal{L}$. However, this approach turns out to suffer from nontrivial size distortion as shown in our simulation study below. To address this issue, we need to account for the effect of selection and adjust the critical value accordingly. The remainder of this subsection is devoted to this task.

More notation is needed. Let $\mathbf{I}_n$ denote the n-dimensional identity matrix, $\boldsymbol{\epsilon} \equiv (\epsilon_t)_{1 \le t \le n}$, and $\mathbf{G} \equiv (g(X_t))_{1 \le t \le n}$. By convention, all vectors are column vectors. For any $\mathcal{S} \subseteq \mathcal{M}$, denote $\mathbf{P}_{\mathcal{S}} \equiv (P_{\mathcal{S}}(X_1), \ldots, P_{\mathcal{S}}(X_n))^{\top}$. When $\mathcal{S} = \mathcal{M}$, we suppress its subscript by simply writing $\mathbf{P} = \mathbf{P}_{\mathcal{M}}$. In addition, let $\widetilde{\mathbf{P}}_{\mathcal{S}}$ denote the residual matrix obtained from projecting $\mathbf{P}_{\mathcal{S}}$ onto $\mathbf{P}_{\mathcal{M}_0}$. That is, $\widetilde{\mathbf{P}}_{\mathcal{S}} \equiv \mathbf{D}_n \mathbf{P}_{\mathcal{S}}$, where $\mathbf{D}_n \equiv \mathbf{I}_n - \mathbf{P}_{\mathcal{M}_0}(\mathbf{P}_{\mathcal{M}_0}^{\top}\mathbf{P}_{\mathcal{M}_0})^{-1}\mathbf{P}_{\mathcal{M}_0}^{\top}$. Finally, we define $\hat{\mathbf{s}}$ as a $|\mathcal{L} \setminus \mathcal{M}_0|$-dimensional vector that collects the signs of $\hat{\beta}_j^{Lasso}$ for $j \in \{j : \hat{\beta}_j^{Lasso} \neq 0\} \setminus \mathcal{M}_0$.

We first characterize the selection event. By the Karush–Kuhn–Tucker conditions for the Lasso problem (2.8), the selection event can be represented by a system of linear-inequality restrictions on $n^{-1/2}\mathbf{P}^{\top}\boldsymbol{\epsilon} = n^{-1/2}\sum_{t=1}^{n} P(X_t)\,\epsilon_t$, which is the score vector of the series regression using the entire dictionary of regressors.[5] Specifically, for any nonrandom selection $\mathcal{S}$ satisfying $\mathcal{M}_0 \subseteq \mathcal{S} \subseteq \mathcal{M}$ and a sign vector $\mathbf{s} \in \{\pm 1\}^{|\mathcal{S} \setminus \mathcal{M}_0|}$, we have

$$\{\mathcal{L} = \mathcal{S}, \hat{\mathbf{s}} = \mathbf{s}\} = \left\{ n^{-1/2}\mathbf{P}^{\top}\boldsymbol{\epsilon} \in \Pi(\mathcal{S}, \mathbf{s}, \lambda_n) \right\}. \tag{2.13}$$

Here, $\Pi(\mathcal{S}, \mathbf{s}, \lambda_n)$ is an m-dimensional (random) polytope given by

$$\Pi(\mathcal{S}, \mathbf{s}, \lambda_n) \equiv \left\{ z \in \mathbb{R}^m : \begin{array}{c} \operatorname{diag}(\mathbf{s})\,(A_{\mathcal{S}} z + c_{\mathcal{S}}) > n^{-1/2}\lambda_n b_{\mathcal{S}}(\mathbf{s}) \quad \text{and} \\ n^{-1/2}\lambda_n b'_{l,\mathcal{S}}(\mathbf{s}) < A'_{\mathcal{S}} z + c'_{\mathcal{S}} < n^{-1/2}\lambda_n b'_{u,\mathcal{S}}(\mathbf{s}) \end{array} \right\}, \tag{2.14}$$

where $\operatorname{diag}(\mathbf{s})$ is a diagonal matrix with its diagonal components given by $\mathbf{s}$,

$$\left\{ \begin{array}{lcl} c_{\mathcal{S}} & \equiv & n^{1/2}\left(\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^{\top}\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}\right)^{-1}\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^{\top}\mathbf{D}_n\mathbf{G}, \\ c'_{\mathcal{S}} & \equiv & n^{-1/2}\widetilde{\mathbf{P}}_{\mathcal{M}\setminus\mathcal{S}}^{\top}\left(\mathbf{I}_n - \widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}\left(\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^{\top}\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}\right)^{-1}\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^{\top}\right)\mathbf{D}_n\mathbf{G}, \end{array} \right. \tag{2.15}$$

[5] See Lemma SA.1 in the Supplemental Appendix for details, which extends a similar result in Lee, Sun, Sun, and Taylor (2016) by allowing for the prior choice set $\mathcal{M}_0$ and penalty weights ω_j.

and $b_{\mathcal{S}}(\mathbf{s})$, $b'_{l,\mathcal{S}}(\mathbf{s})$, $b'_{u,\mathcal{S}}(\mathbf{s})$, $A_{\mathcal{S}}$, and $A'_{\mathcal{S}}$ are directly observable quantities. The latter observable quantities do not pose any difficulty in our theoretical analysis (though they are needed for implementation). We thus defer their somewhat complicated definitions to Appendix A.1 to streamline the discussion; see (A.1). On the contrary, the random vectors $c_{\mathcal{S}}$ and $c'_{\mathcal{S}}$ are unobservable because $\mathbf{G}$ involves the unknown $g(\cdot)$ function. For this reason, the structure of the polytope $\Pi(\mathcal{S}, \mathbf{s}, \lambda_n)$ is not directly observed, either.

The significance of the above (non-asymptotic) characterization is that it precisely depicts the relation between the Lasso-assisted selection and the subsequent series estimation in finite samples, through their common dependence on the score vector $n^{-1/2}\mathbf{P}^{\top}\boldsymbol{\epsilon}$. To see this more clearly, recall that for any given selection $\mathcal{S}$, the asymptotic normality (formulated in terms of strong Gaussian coupling) of the $\hat{g}_{\mathcal{S}}(\cdot)$ estimator is driven by the score $n^{-1/2}\mathbf{P}_{\mathcal{S}}^{\top}\boldsymbol{\epsilon}$, which is a subvector of $n^{-1/2}\mathbf{P}^{\top}\boldsymbol{\epsilon}$. However, when $\mathcal{S}$ is selected by Lasso with a particular sign configuration $\mathbf{s}$, the score $n^{-1/2}\mathbf{P}^{\top}\boldsymbol{\epsilon}$ is restricted within the polytope $\Pi(\mathcal{S}, \mathbf{s}, \lambda_n)$. This restriction would modify the score's asymptotic normality into a form of truncated normality. Roughly speaking, the data-driven selection may make the originally exogenous conditioning variables "effectively endogenous" through truncating the support of the score. A failure to account for this effect would generally lead to size distortion. It is important to note that this type of size distortion is distinct from the usual small-sample phenomenon that central limit theorems may not "kick in" sufficiently well in a moderately sized sample; indeed, the aforementioned truncation effect would arise even if the score is exactly normally distributed (say, in a Gaussian model with fixed design).

We now propose a new critical value to adjust for the truncation effect. The key is to construct a feasible approximation for the unobserved polytope $\Pi(\mathcal{S}, \mathbf{s}, \lambda_n)$ for any given $\mathcal{S}$ that contains $\mathcal{M}_0$. As mentioned above, the polytope is not directly observed because $\mathbf{G}$ is unknown. To construct an approximation for $\mathbf{G}$, we regress $\mathbf{Y}$ on $\mathbf{P}_{\mathcal{S}}$ with the resulting regression coefficient given by

$$\widehat{b}_{\mathcal{S}} \equiv \left(\mathbf{P}_{\mathcal{S}}^{\top}\mathbf{P}_{\mathcal{S}}\right)^{-1}\mathbf{P}_{\mathcal{S}}^{\top}\mathbf{Y}. \tag{2.16}$$

We then apply a truncation on this least-squares estimator to obtain $\tilde{\beta}_{\mathcal{S}}$, with its jth component given by

$$\tilde{\beta}_{\mathcal{S},j} \equiv \widehat{b}_{\mathcal{S},j} \cdot 1\left\{|\widehat{b}_{\mathcal{S},j}| \geq \log(n) n^{-1/2}\widehat{\sigma}_{\mathcal{S},j}\right\}, \tag{2.17}$$

where $\widehat{b}_{\mathcal{S},j}$ denotes the jth component of $\widehat{b}_{\mathcal{S}}$, and $\widehat{\sigma}_{\mathcal{S},j}$ is the estimated standard error of $\widehat{b}_{\mathcal{S},j}$ obtained as the square-root of the jth diagonal element of $\widehat{Q}_{\mathcal{S}}^{-1}\widehat{\Sigma}_{\mathcal{S}}\widehat{Q}_{\mathcal{S}}^{-1}$.[6] The n-dimensional vector

[6] The intuition for using the truncation is as follows. If the estimator $\widehat{b}_{\mathcal{S},j}$ corresponds to a zero coefficient in the population, $\widehat{b}_{\mathcal{S},j}/(n^{-1/2}\widehat{\sigma}_{\mathcal{S},j})$ is approximately $\mathcal{N}(0,1)$. In addition, these "zero" t-statistics are uniformly bounded by the $\log(n)$ factor with probability approaching 1. The truncation shrinks these noisy estimates of zero directly

$\mathbf{G}$ is then approximated by the (truncated) projection $\mathbf{P}_{\mathcal{S}}\tilde{\beta}_{\mathcal{S}}$. Plugging this approximation into (2.15), we further obtain (after simplifying the expressions) approximations for $c_{\mathcal{S}}$ and $c'_{\mathcal{S}}$ in the form of

$$\widehat{c}_{\mathcal{S}} = n^{1/2}\tilde{\beta}_{\mathcal{S}\setminus\mathcal{M}_0}, \quad \widehat{c}'_{\mathcal{S}} = 0,$$

where $\tilde{\beta}_{\mathcal{S}\setminus\mathcal{M}_0}$ is the subvector of $\tilde{\beta}_{\mathcal{S}}$ extracted in accordance with $\mathcal{S}\setminus\mathcal{M}_0$ as a subset of $\mathcal{S}$. A feasible proxy for $\Pi(\mathcal{S}, \mathbf{s}, \lambda_n)$ can then be obtained by replacing $(c_{\mathcal{S}}, c'_{\mathcal{S}})$ with $(\widehat{c}_{\mathcal{S}}, \widehat{c}'_{\mathcal{S}})$ defined in (2.14), that is,

$$\widehat{\Pi}(\mathcal{S}, \mathbf{s}, \lambda_n) \equiv \left\{ z \in \mathbb{R}^m : \begin{array}{c} \operatorname{diag}(\mathbf{s})\,(A_{\mathcal{S}}z + n^{1/2}\tilde{\beta}_{\mathcal{S}\setminus\mathcal{M}_0}) > n^{-1/2}\lambda_n b_{\mathcal{S}}(\mathbf{s}) \;\text{ and} \\ n^{-1/2}\lambda_n b'_{l,\mathcal{S}}(\mathbf{s}) < A'_{\mathcal{S}}z < n^{-1/2}\lambda_n b'_{u,\mathcal{S}}(\mathbf{s}) \end{array} \right\}. \tag{2.18}$$

We are now ready to construct the new critical value. Let $\widetilde{N}^*$ be an m-dimensional standard Gaussian random vector that is independent of the data. For a given selection $\mathcal{S}$, define $\widetilde{N}^*_{\mathcal{S}}$ as the subvector of $\widehat{\Sigma}^{1/2}_{\mathcal{M}}\widetilde{N}^*$ extracted in accordance with $\mathcal{S}$ as a subset of $\mathcal{M}$, and use it to compute $\widetilde{T}^*_{\mathcal{S}}$ as described in (2.12). We then set

$$cv_{\mathcal{S},\alpha} \equiv \inf\left\{ u \in \mathbb{R} : \frac{\mathbb{P}^*\left(\widetilde{T}^*_{\mathcal{S}} \geq u, \widehat{\Sigma}^{1/2}_{\mathcal{M}}\widetilde{N}^* \in \widehat{\Pi}(\mathcal{S}, \mathbf{s}, \lambda_n)\right)}{\mathbb{P}^*\left(\widehat{\Sigma}^{1/2}_{\mathcal{M}}\widetilde{N}^* \in \widehat{\Pi}(\mathcal{S}, \mathbf{s}, \lambda_n)\right)} = \alpha \right\}, \tag{2.19}$$

where $\mathbb{P}^*(\cdot)$ denotes the conditional distribution of $\widetilde{N}^*$ given data.[7] Our proposed critical value is obtained by evaluating $cv_{\mathcal{S},\alpha}$ at $\mathcal{S} = \mathcal{L}$, that is,

$$cv_{\mathcal{L},\alpha} \equiv cv_{\mathcal{S},\alpha}\big|_{\mathcal{S}=\mathcal{L}}. \tag{2.20}$$

The selective test rejects the null hypothesis in (2.1) if $\widehat{T}_{\mathcal{L}} > cv_{\mathcal{L},\alpha}$.

The intuition for the proposed critical value is as follows.[8] Note that the (conditionally) Gaussian vector $\widehat{\Sigma}^{1/2}_{\mathcal{M}}\widetilde{N}^*$ provides a distributional approximation for the score vector $n^{-1/2}\mathbf{P}^{\top}\boldsymbol{\epsilon}$. Since $\widetilde{T}^*_{\mathcal{S}}$ is formed using the subvector $\widetilde{N}^*_{\mathcal{S}}$, $\widehat{\Sigma}^{1/2}_{\mathcal{M}}\widetilde{N}^*$ and $\widetilde{T}^*_{\mathcal{S}}$ provide a joint distributional approximation for the score $n^{-1/2}\mathbf{P}^{\top}\boldsymbol{\epsilon}$ and the sup-t statistic $\widehat{T}_{\mathcal{S}}$ under the null hypothesis. As such, the joint asymptotic behavior of the test statistic and the selection event $\{n^{-1/2}\mathbf{P}^{\top}\boldsymbol{\epsilon} \in \Pi(\mathcal{S}, \mathbf{s}, \lambda_n)\}$ is captured by that of $\widetilde{T}^*_{\mathcal{S}}$ and $\{\widehat{\Sigma}^{1/2}_{\mathcal{M}}\widetilde{N}^* \in \widehat{\Pi}(\mathcal{S}, \mathbf{s}, \lambda_n)\}$. The critical value described in (2.19) is simply

to zero. This noise-reduction generally leads to better performance in finite samples.

[7] This critical value may be computed by simulating the Gaussian random vector $\widetilde{N}^*$. A computationally more efficient method is to sample directly from the truncated normal distribution in restriction to the selection event. The Matlab package accompanying this paper follows the latter computational strategy by using the minimax tilting algorithm proposed in Botev (2016).

[8] Appendix A.2 provides a pedagogical example, where we use a simple linear regression model to illustrate more concretely the effect of model selection on the subsequent inference and the intuition for the new critical value.

defined as a tail quantile of the conditional distribution of $\widetilde{T}_{\mathcal{S}}^{*}$ in restriction to the "coupling" selection event $\{\widehat{\Sigma}_{\mathcal{M}}^{1/2}\widetilde{N}^{*} \in \widehat{\Pi}(\mathcal{S}, \mathbf{s}, \lambda_n)\}$, which captures how the polytope restriction on the score vector distorts the distribution of the sup-t statistic.[9]

We close this subsection with a couple of remarks. We first note that our strategy for correcting the critical value is not restricted to the Lasso method. For the other methods such as the group Lasso (Yuan and Lin (2006)) and the elastic net (Zou and Hastie (2005), Zou and Zhang (2009)), one may modify the underlying Karush–Kuhn–Tucker conditions accordingly and characterize the selection event in a similar fashion as (2.13). Critical values may then be constructed from the corresponding conditional coupling distributions. Secondly, we stress that our analysis focuses on testing whether $g(\cdot) = 0$, and hence, our inference concentrates on the null hypothesis. Another open question is how to make uniform inference for the unknown function $g(\cdot)$ also under the alternative, while properly accounting for the selection effect. The latter question is more challenging because, under "local" alternatives, the selection may miss "moderate" features of $g(\cdot)$, and lead to non-negligible biases for inference.[10] This is not an issue (in terms of size control) for our hypothesis-testing problem because under the null $g(\cdot)$ is known to be zero.

2.3 Asymptotic properties of the selective test

We now establish the asymptotic properties of the selective test. We shall show that the proposed test has valid size control under the null hypothesis. We also analyze the test's power under local alternatives so as to theoretically clarify how the Lasso-assisted selection helps improve power. In this subsection, we focus on the baseline setting in which Y_t is directly observed. The result can be straightforwardly extended to allow Y_t to depend on some unknown parameter θ^*; see Section 3 below. We start with introducing a few regularity conditions.

[9] One may wonder whether the size correction can be automatically achieved via resampling methods such as the bootstrap. We investigate this possibility through a simulation study in Section SC.2 of the online supplemental appendix. The simulation results show that a test based on the i.i.d. bootstrap tends to be very conservative and have poor power (even if there is no serial dependence in the data). The theoretical investigation of resampling methods for the selective test is beyond the scope of this paper and is left for future research.

[10] The uniform nonparametric inference with a data-driven selection of series terms should be distinguished from a recent strand of literature on "selective inference." For example, in a Gaussian linear model, Lee, Sun, Sun, and Taylor (2016) study the inference for the coefficients of a submodel selected by Lasso. That research question is very different from making uniform nonparametric inference, because it would effectively shift the inferential target from $g(\cdot)$ to a statistically selected submodel; the latter is a "moving target" that could be very different from the original object of interest. That being said, selective inference may be fruitfully used in many other econometric problems, as demonstrated in the recent interesting work by Cox and Shi (2019), Liao and Shi (2020), Andrews, Kitagawa, and McCloskey (2021a,b); our coupling-based inference technique might be useful to extend that line of research to growing-dimensional or functional settings for general serially dependent data.

Assumption 1. *(i) The eigenvalues of* $Q_{\mathcal{M}}$ *and* $\Sigma_{\mathcal{M}}$ *are bounded from above and away from zero; (ii) there exists a sequence of* m*-dimensional standard Gaussian random vectors* $\widetilde{N}_n$ *such that*

$$n^{-1/2}\sum_{t=1}^{n} P\left(X_t\right)\epsilon_t = \Sigma_{\mathcal{M}}^{1/2}\widetilde{N}_n + o_p(\log(n)^{-1});$$

(iii) $\|\widehat{Q}_{\mathcal{M}} - Q_{\mathcal{M}}\| + \|\widehat{\Sigma}_{\mathcal{M}} - \Sigma_{\mathcal{M}}\| = o_p((m^{1/2}\log(n))^{-1})$*; (iv)* $m = o(n)$ *and* $\log(\zeta_n^L) = O(\log(m))$*, where* $\zeta_n^L \equiv \sup_{x_1,x_2\in\mathcal{X}} \|P(x_1) - P(x_2)\| / \|x_1 - x_2\|$*.*

Assumption 1 imposes high-level conditions that are similar to those used in prior work on uniform series-based inference. Condition (i) is fairly standard for series estimation; see, for example, Andrews (1991a), Newey (1997), and Chen (2007). Condition (ii) requires that the scaled score sequence $n^{-1/2}\sum_{t=1}^{n} P(X_t)\epsilon_t$ admits a Gaussian coupling in the growing-dimensional setting (i.e., $m \to \infty$), which may be verified by applying Yurinskii's coupling for i.i.d. data, or the theory of Li and Liao (2020) in the more general time-series setting for heterogeneous mixingales. This condition is crucial for our purpose of making uniform functional inference and it also implicitly imposes the binding restriction within our analysis on how fast m may grow with the sample size (especially in the presence of series dependence); see Li and Liao (2020) for a more detailed technical discussion.[11] Condition (iii) pertains to the convergence rates of $\widehat{Q}_{\mathcal{M}}$ and $\widehat{\Sigma}_{\mathcal{M}}$, which can be verified under primitive conditions as shown in Chen and Christensen (2015) and Li and Liao (2020). Condition (iv) is trivially satisfied by commonly used series basis.

To set the stage for the local power analysis, we consider a sequence of data generating processes under which $\mathbb{E}\left[Y_t|X_t = x\right] = g_n(x)$, where $g_n(\cdot)$ is a (possibly) drifting sequence of functions. These functions are assumed to satisfy the following.

Assumption 2. *(i) There exists a sequence* $(b_n^*)_{n\geq 1}$ *of* m*-dimensional constant vectors such that*

$$\sup_{x\in\mathcal{X}} n^{1/2}\left|g_n(x) - P(x)^{\top}b_n^*\right| = O(1);$$

(ii) there exists a subset $\mathcal{R} \subseteq \mathcal{M}_0^c$ *such that* $\min_{j\in\mathcal{R}}|b_{n,j}^*| > 0$ *and* $b_{n,j}^* = 0$ *when* $j \in \mathcal{M}_0^c \setminus \mathcal{R}$*.*

Assumption 2(i) states that the $g_n(\cdot)$ function may be approximately represented by the growing-dimensional b_n^* vector, which specifies how $g_n(\cdot)$ loads on the basis functions. This is

[11] If one restricts attention to the setting with i.i.d. data, it might be possible to generalize our result to allow m to grow faster possibly under some additional sparsity restriction (which is not assumed here). We do not pursue that extension in the present paper because our primary goal is to accommodate serial dependence commonly seen in time-series settings so as to facilitate macro and finance applications, and certain applied-micro applications involving panel data. Establishing a theory for $m > n$ under sparsity is not our main focus, but might be an interesting topic for future research; this could be technically challenging in a setting with general serially dependent data (e.g., mixingales) as considered here.

well understood in series estimation, for which comprehensive results are available from the literature on numerical approximation (see, e.g., Chen (2007)); this setup also directly accommodates linear specifications with "many regressors." Given this representation, condition (ii) further introduces a "relevance set," $\mathcal{R}$, which marks all basis functions in $\mathcal{M}_0^c$ (on which the Lasso selection is active) with nonzero loadings. Note that $\mathcal{R}$ is empty under the null hypothesis, but it plays an important role under the alternative.

Intuitively, if the user knew the (actually unknown) structure of $\mathcal{R}$ a priori, it would be natural to combine it with their prior choice $\mathcal{M}_0$ to form an "oracle" selection

$$\mathcal{M}^\star \equiv \mathcal{M}_0 \cup \mathcal{R},$$

which is arguably the best one may wish to obtain from any selection algorithm (e.g., Lasso). The $\mathcal{M}^\star$ set thus depicts the intrinsic complexity of $g_n(\cdot)$ given the user's ex ante choice (including the dictionary $\mathcal{M}$ and the prior choice $\mathcal{M}_0$). In this sense, $g_n(\cdot)$ is the most complex when $\mathcal{M}^\star = \mathcal{M}$, because one would use all basis functions to conduct the series estimation. On the other extreme, if $\mathcal{M}^\star$ is "sparse" in the sense that it contains only a few elements, $g_n(\cdot)$ is "effectively parametric," and hence, relatively simple to uncover. Consistent with this logic, our theory presented below shows that the selective test satisfies an *adaptive* property, namely, it is more powerful when the alternative is less complex. In our analysis, it turns out that the aforementioned notion of complexity may be more precisely quantified as (with $\|\cdot\|$ denoting the Euclidean norm)

$$\kappa(\mathcal{M}^\star) \equiv \sup_{x\in\mathcal{X}} \|P_{\mathcal{M}^\star}(x)\|,$$

which is a non-decreasing function of $\mathcal{M}^\star$ with respect to the partial order of set inclusion. Hence, a larger $\mathcal{M}^\star$ corresponds to a higher value or faster divergence rate of $\kappa(\mathcal{M}^\star)$, and vice versa.[12]

We also need the following condition on the Lasso penalty. For any $m_1 \times m_2$ real matrix $A = [A_{ij}]_{1\le i\le m_1, 1\le j\le m_2}$, we denote $\|A\|_1 \equiv \max_{1\le j\le m_2}\sum_{i=1}^{m_1}|A_{ij}|$.

Assumption 3. *The penalty parameters λ_n and $\{\omega_j\}_{j\in\mathcal{M}_0^c}$ satisfy*

$$\frac{(n\log(m))^{1/2}}{\lambda_n \min_{j\in\mathcal{M}_0^c\setminus\mathcal{R}}\omega_j} + \frac{|\mathcal{R}|^{1/2}n^{-1/2}\lambda_n \max_{j\in\mathcal{R}}\omega_j + \log(n)}{n^{1/2}\min_{j\in\mathcal{R}}|b^*_{n,j}|} = o_p(1) \tag{2.21}$$

and, for some fixed $\eta \in (0,1)$,

$$\frac{\max_{j\in\mathcal{R}}\omega_j}{\min_{j\in\mathcal{M}_0^c\setminus\mathcal{R}}\omega_j}\left\|(\widetilde{\mathbf{P}}_{\mathcal{R}}^\top\widetilde{\mathbf{P}}_{\mathcal{R}})^{-1}\widetilde{\mathbf{P}}_{\mathcal{R}}^\top\widetilde{\mathbf{P}}_{\mathcal{M}_0^c\setminus\mathcal{R}}\right\|_1 \le 1-\eta \tag{2.22}$$

with probability approaching 1.

[12] In our theory, $\mathcal{M}^\star$ is allowed to contain a growing number of elements. Therefore, $\kappa(\mathcal{M}^\star)$ is typically a divergent sequence of positive numbers and its "magnitude" is gauged by its growth rate to infinity.

Assumption 3 mainly ensures that the Lasso estimator described in (2.8) is sign-consistent under the null and alternative hypotheses.[13] This condition is high-level in nature and it does not directly pin down any specific penalty scheme. In Appendix A.1, we provide a concrete feasible choice that fulfills this technical condition.

We are now ready to state the asymptotic size and power properties of the selective test, which is the main result of this paper. Below, for two sequences of positive numbers a_n and b_n, we write $a_n \succ b_n$ if $a_n \geq c_n b_n$ for some strictly positive sequence $c_n \to \infty$.

Theorem 1. *Under Assumptions 1, 2, and 3, the following statements hold for any significance level $\alpha \in (0, 1/2)$: (a) The selective test has asymptotic level α under the null hypothesis (2.1), that is, $\mathbb{P}(\widehat{T}_{\mathcal{L}} > cv_{\mathcal{L},\alpha}) \to \alpha$; (b) the selective test is consistent against any local alternative that satisfies*

$$\sup_{x \in \mathcal{X}} |g_n(x)| \succ \kappa\left(\mathcal{M}^\star\right) \log(n)^{1/2} n^{-1/2}, \tag{2.23}$$

that is, $\mathbb{P}(\widehat{T}_{\mathcal{L}} > cv_{\mathcal{L},\alpha}) \to 1$.

Part (a) of Theorem 1 shows that the selective test has valid size control under the null hypothesis. Part (b) further establishes the consistency of the test against local alternatives that satisfy condition (2.23), with the "boundary" of the local neighborhood (under the uniform metric) characterized by the $\kappa\left(\mathcal{M}^\star\right) \log(n)^{1/2} n^{-1/2}$ rate.

The local power result deserves some additional discussion. Its key significance is to provide a sense in which the selective test is adaptive with respect to the complexity of $g_n(\cdot)$ as gauged by $\kappa\left(\mathcal{M}^\star\right)$. That is, the test is able to consistently detect a faster-vanishing nonzero sequence of $\sup_{x \in \mathcal{X}} |g_n(x)|$ when the $g_n(\cdot)$ function is easier to approximate (i.e., $\mathcal{M}^\star$ is smaller), despite the fact that this information is unknown a priori. This is an important improvement relative to the benchmark non-selective method (cf. Belloni, Chernozhukov, Chetverikov, and Kato (2015) and Li and Liao (2020)). Indeed, the non-selective method employs the entire dictionary $\mathcal{M}$ of basis functions, which can be considered as a corner case of the selective test corresponding to the most conservative prior choice $\mathcal{M}_0 = \mathcal{M}$. The power of the non-selective test is thus always dictated by the fast-diverging sequence $\kappa\left(\mathcal{M}\right)$, and hence low, regardless of the actual complexity underlying the data generating process (i.e., it is non-adaptive).

To further appreciate the adaptiveness of the selective test, we consider another "corner" case in which the set $\mathcal{M}^\star$ only contains a bounded number of elements. This corresponds to a situation

[13] Recall that $\mathcal{R}$ is empty under the null hypothesis. By convention, we set the maximum and minimum of a collection of nonnegative numbers over the empty set to 0 and ∞, respectively. Under this convention, Assumption 3 can be reduced to a simpler form under the null hypothesis, that is, $\sqrt{n \log(m)}/(\lambda_n \min_{j \in \mathcal{M}_0^c} \omega_j) = o_p(1)$.

in which $g_n(\cdot)$ under the alternative has a parametric but a priori unknown form. In this case, $\kappa(\mathcal{M}^\star)$ is bounded and the $\kappa(\mathcal{M}^\star)\log(n)^{1/2}n^{-1/2}$ rate can be simplified as $\log(n)^{1/2}n^{-1/2}$, which is essentially the $n^{-1/2}$ parametric rate. Attaining this nearly parametric rate is remarkable because the user was "prepared" to conduct a nonparametric analysis, as they did not know beforehand that $g_n(\cdot)$ has a simple form, let alone its specific parametric specification among a large number (i.e., 2^m with $m \to \infty$) of possibilities. In sharp contrast, the non-selective method has power only at the well-known and much slower nonparametric rate.

3 Extensions

We consider two extensions for our baseline method developed in the previous section. Section 3.1 presents a different version of the selective test based on an alternative test statistic. Section 3.2 describes two approaches for handling unknown finite-dimensional parameters.

3.1 Selective test with an alternative test statistic

The $\widehat{T}_{\mathcal{L}}$ test statistic is based on the supremum of the (studentized) series estimator for $g(\cdot)$, which quantifies deviations from the null under the uniform metric on the functional space. This is a natural choice in the nonparametric setting, as $g(\cdot)$ is the model primitive in that context. That noted, the underlying econometric idea can be easily extended to accommodate the other types of test statistics. In this subsection, we provide a concrete example to demonstrate how our baseline theory may be modified for that purpose.

Consider an alternative test statistic defined as the maximum of the t-statistics associated with individual regression coefficients in the series regression. Since this statistic is directly based on the regression coefficients, it is perhaps better suited than $\widehat{T}_{\mathcal{L}}$ for studying linear models with "many" regressors. Specifically, for each given selection $\mathcal{S}$, we define the test statistic as

$$\widehat{T}'_{\mathcal{S}} \equiv \max_{1\leq j\leq|\mathcal{S}|} \frac{n^{1/2}\big|\widehat{b}_{\mathcal{S},j}\big|}{\widehat{\sigma}_{\mathcal{S},j}}, \tag{3.24}$$

where we recall that $\widehat{b}_{\mathcal{S},j}$ is the jth component of the series regression coefficient $\widehat{b}_{\mathcal{S}}$ (see (2.16)) and $\widehat{\sigma}_{\mathcal{S},j}$ is the estimated standard error obtained as the square-root of the jth diagonal element of $\widehat{Q}_{\mathcal{S}}^{-1}\widehat{\Sigma}_{\mathcal{S}}\widehat{Q}_{\mathcal{S}}^{-1}$. Its feasible distributional "coupling" is given by

$$\widetilde{T}'^{*}_{\mathcal{S}} \equiv \max_{1\leq j\leq|\mathcal{S}|} \frac{\big|\big[\widehat{Q}_{\mathcal{S}}^{-1}\widetilde{N}^{*}_{\mathcal{S}}\big]_j\big|}{\widehat{\sigma}_{\mathcal{S},j}}, \tag{3.25}$$

where the $[\,\cdot\,]_j$ operator extracts the jth component of a vector and $\widetilde{N}^{*}_{\mathcal{S}}$ is a subvector exacted from $\widehat{\Sigma}_{\mathcal{M}}^{1/2}\widetilde{N}^{*}$ in accordance with $\mathcal{S}$ as a subset of $\mathcal{M}$ for some generic m-dimensional standard normal random vector $\widetilde{N}^*$. Analogous to (2.19), we define the critical value for the new test statistic as

$$cv'_{\mathcal{S},\alpha} \equiv \inf \left\{ u \in \mathbb{R} : \frac{\mathbb{P}^* \left(\widetilde{T}'^*_{\mathcal{S}} \geq u, \widehat{\Sigma}^{1/2}_{\mathcal{M}} \widetilde{N}^* \in \widehat{\Pi}(\mathcal{S}, \mathbf{s}, \lambda_n) \right)}{\mathbb{P}^* \left(\widehat{\Sigma}^{1/2}_{\mathcal{M}} \widetilde{N}^* \in \widehat{\Pi}(\mathcal{S}, \mathbf{s}, \lambda_n) \right)} = \alpha \right\}. \tag{3.26}$$

Finally, we set

$$\widehat{T}'_{\mathcal{L}} = \widehat{T}'_{\mathcal{S}} \big|_{\mathcal{S}=\mathcal{L}}, \quad cv'_{\mathcal{L},\alpha} \equiv cv'_{\mathcal{S},\alpha} \big|_{\mathcal{S}=\mathcal{L}}. \tag{3.27}$$

This alternative "sup-b" selective test rejects the null hypothesis in (2.1) if $\widehat{T}'_{\mathcal{L}} > cv'_{\mathcal{L},\alpha}$. Similar to Theorem 1, we have the following result for its asymptotic properties.

Theorem 2. *Under Assumptions 1, 2, and 3, the following statements hold for any significance level $\alpha \in (0, 1/2)$: (a) The sup-b selective test has asymptotic level α under the null hypothesis (2.1), that is, $\mathbb{P}(\widehat{T}'_{\mathcal{L}} > cv'_{\mathcal{L},\alpha}) \to \alpha$; (b) the sup-b selective test is consistent against any local alternative that satisfies*

$$\max_{1 \leq j \leq |\mathcal{M}^\star|} \left| b^*_{n,j} \right| \succ \log(n)^{1/2} n^{-1/2}, \tag{3.28}$$

that is, $\mathbb{P}(\widehat{T}'_{\mathcal{L}} > cv'_{\mathcal{L},\alpha}) \to 1$.

3.2 The case with unknown parameters

So far, we have analyzed the selective test in the baseline setting in which Y_t is directly observable. As the examples in Section 2.1 show, Y_t may depend on an unknown parameter θ^* in some empirical applications. In this subsection, we describe how the selective test may be applied in this more general setting. Below, we write $Y_t(\theta)$ to emphasize the dependence of Y_t on a generic parameter value $\theta \in \Theta$ and, correspondingly, use $\widehat{T}_{\mathcal{L}}(\theta)$ and $cv_{\mathcal{L},\alpha}(\theta)$ to denote the selective test statistic and the critical value (recall (2.10) and (2.20)) computed using $Y_t = Y_t(\theta)$. The null hypothesis of interest concerns the conditional moment restriction evaluated at the true value θ^*, that is, $H_0 : g(\cdot) = 0$, where $g(x) \equiv \mathbb{E}[Y_t(\theta^*)|X_t = x]$.

Arguably the most straightforward approach for dealing with the unknown parameter is to construct the Anderson–Rubin confidence set by inverting the selective test. Specifically, for each candidate parameter value $\theta \in \Theta$, we implement the selective test for the null hypothesis

$$H_{0,\theta} : \mathbb{E}[Y_t(\theta)|X_t = x] = 0 \text{ for all } x \in \mathcal{X}.$$

The $1-\alpha$ level Anderson–Rubin confidence set for the true value θ^* is then constructed as

$$CS_{1-\alpha} \equiv \left\{ \theta \in \Theta : \widehat{T}_{\mathcal{L}}(\theta) \leq cv_{\mathcal{L},\alpha}(\theta) \right\},$$

which collects the θ's such that the selective test does not reject. By the duality between test and confidence set, Theorem 1 implies that $\theta^* \in CS_{1-\alpha}$ with probability approaching $1-\alpha$. We reject the original null hypothesis (i.e., $g(\cdot) = 0$) when the confidence set $CS_{1-\alpha}$ is empty.

The Anderson–Rubin approach has a well-known desirable feature that it is robust against the weak/partial identification of the unknown parameter θ^*. This issue is particularly relevant for the empirical analysis of macro-style models (see, e.g., Stock and Wright (2000)). Although making this type of robust inference on θ^* is not our primary goal, it is a "free" by-product of the proposed test.[14]

The downside of the Anderson–Rubin approach, however, is that inverting the test for a large number of candidate values may be computationally expensive. For this reason, we also consider a more practical "plug-in" approach. Suppose that an estimate for θ^*, denoted $\hat{\theta}$, is available. We assume that $\hat{\theta}$ is $n^{1/2}$-consistent for θ^* but do not impose any additional specific structure on it. This agnostic setup is intentionally designed to accommodate applications in which $\hat{\theta}$ is calibrated (possibly in other studies based on external datasets that are unavailable), which is quite typical in macro-style applications (see Chodorow-Reich and Karabarbounis (2016) for interesting examples). Below, we propose a set of conditions under which the $O_p(n^{-1/2})$ estimation error in $\hat{\theta}$ is asymptotically negligible for our testing purpose. Given the lack of information regarding $\hat{\theta}$, this is arguably the only reasonable way to proceed. The resulting plug-in selective test rejects the null hypothesis when $\widehat{T}_{\mathcal{L}}(\hat{\theta}) > cv_{\mathcal{L},\alpha}(\hat{\theta})$.

Assumption 4. *(i) The estimator $\hat{\theta}$ satisfies $n^{1/2}(\hat{\theta}-\theta^*) = O_p(1)$; (ii) $Y_t(\theta)$ is twice continuously differentiable in θ with bounded derivatives; (iii) $n^{-1}\sum_{t=1}^{n} P(X_t)(\partial_\theta Y_t(\theta^*))^\top - \Gamma = o_p(\log(n)^{-1})$, where $\Gamma \equiv n^{-1}\sum_{t=1}^{n} \mathbb{E}\left[P(X_t)(\partial_\theta Y_t(\theta^*))^\top\right]$; (iv) the function $h_n(x) \equiv \mathbb{E}\left[\partial_\theta Y_t(\theta^*) \middle| X_t = x\right]$ does not depend on t, and there exist some constant $r \geq 1/2$ and a matrix-valued sequence ϕ_n^* such that $\sup_{x\in\mathcal{X}} \|\phi_n^* P_{\mathcal{M}_0}(x) - h_n(x)\| = O(|\mathcal{M}_0|^{-r})$; (v) for some constant $C > 0$, $\inf_{x\in\mathcal{X}} \|P_{\mathcal{M}_0}(x)\| \geq C|\mathcal{M}_0|^{1/2}$ and $|\mathcal{M}_0| \geq C\log(m)^{3/2}$.*

Assumption 4(i) states that $\hat{\theta}$ is a $n^{1/2}$-consistent estimator for θ^*, which is satisfied by most commonly used estimators.[15] Conditions (ii)–(iv) ensure that the statistics of interest depend on θ in a smooth manner. Condition (v) mainly requires that the size of $\mathcal{M}_0$ grows at least at the $\log(m)^{3/2}$ rate. Note that this condition is not needed in our baseline setting (recall Theorem 1). Here, we require $\mathcal{M}_0$ to diverge so as to ensure that the post-selection series estimation is

[14] Along this line, it might be interesting to extend the selective test to the setting with conditional moment inequalities (see, e.g., Andrews and Shi (2013), Chernozhukov, Lee, and Rosen (2013), Li, Liao, and Quaedvlieg (2020)). But these extensions are clearly beyond the scope of the current paper, and hence, left for future research.

[15] Under the alternative hypothesis, θ^* is interpreted as the pseudo-true parameter.

at least “moderately nonparametric.” By doing so, the statistical noise in the nonparametric test will dominate the fast-converging estimation error in $\hat{\theta}$, which makes the latter asymptotically negligible for the nonparametric inference.

Proposition 1. *Under Assumptions 1, 2, 3, and 4, the assertions in Theorem 1 hold for the plug-in selective test that rejects the null hypothesis when* $\widehat{T}_{\mathcal{L}}(\hat{\theta}) > cv_{\mathcal{L},\alpha}(\hat{\theta})$.

4 Monte Carlo simulations

We examine the finite-sample performance of the proposed selective test in a Monte Carlo experiment. Section 4.1 presents the setting and Section 4.2 reports the results.

4.1 The setting

We consider a bivariate conditioning variable $X_t = (X_{1,t}, X_{2,t})$ simulated as $X_{j,t} = Z_t + v_{j,t}$ for $j = 1, 2$, where Z_t is an autoregressive process generated by

$$Z_t = \rho Z_{t-1} + (1 - \rho^2)^{1/2} \eta_t,$$

and η_t, $v_{1,t}$, and $v_{2,t}$ are i.i.d. standard normal random shocks. We set $\rho = 0.5$ or 0.8 so that X_t may have different levels of persistence, whereas the variance of Z_t is normalized to unity. The dependent variable Y_t is further generated according to $Y_t = g(X_t) + \epsilon_t$, where

$$g(x) = \frac{\delta \exp(x_1 + x_2)}{1 + \exp(x_1 + x_2)}, \quad \epsilon_t = \exp(Z_t)\epsilon_t^*, \quad \epsilon_t^* \overset{i.i.d.}{\sim} \mathcal{N}(0, 1).$$

The ϵ_t^* shock is independent of the other processes, but the disturbance term ϵ_t in the nonparametric regression features conditional heteroskedasticity. The δ parameter plays a key role in our simulation design. When $\delta = 0$, $g(\cdot) = 0$ identically, so the null hypothesis holds. When $\delta \neq 0$, we are under the alternative hypothesis and the magnitude of δ quantifies how far the alternative deviates from the null. Below, we set $\delta = 0$ for the size analysis, and set $\delta \in \{0.1, 0.2, \ldots, 1\}$ to trace out a test’s power curve. The sample size is set as $n = 150$, 250, or 500. The number of Monte Carlo replications is 10,000.

We examine the finite-sample size and power properties of the proposed selective test at significance level $\alpha = 5\%$. To implement the test, we choose the Lasso penalty parameters according to Algorithm A in Appendix A.1, and then implement the test as described in Section 2.2. The prior choice set $\mathcal{M}_0$ only contains the constant term, which is our recommended default choice. For comparison, we also consider two other tests. The first is the non-selective test of Belloni, Chernozhukov, Chetverikov, and Kato (2015) and Li and Liao (2020), which rejects the null hypothesis when $\widehat{T}_{\mathcal{M}}$ exceeds the $1-\alpha$ quantile of $\widetilde{T}^*_{\mathcal{M}}$ given data; recall the definitions in (2.7) and (2.12). The second is the uncorrected selective test, which rejects the null hypothesis when the selective test statistic $\widehat{T}_{\mathcal{L}}$ exceeds the $1-\alpha$ quantile of $\widetilde{T}^*_{\mathcal{L}} \equiv \widetilde{T}^*_{\mathcal{S}}|_{\mathcal{S}=\mathcal{L}}$ given data (i.e., it does not correct for the truncation effect). For simplicity, we refer to the three tests under consideration as the selective, non-selective, and the uncorrected test, respectively.

We need a collection of basis functions to implement these tests. A natural choice is polynomials. Clearly, "plain" polynomial terms of conditioning variables tend to be highly correlated in a given sample. This may lead to numerical instability in the series estimation (e.g., the $n^{-1}\mathbf{P}^\top\mathbf{P}$ matrix may not be inverted with enough numerical precision using commonly used software) especially when a large number of series terms are involved. Following prior work (see, e.g., Li, Liao, and Quaedvlieg (2020)), we mitigate this numerical issue by using the Legendre polynomial. Recall that the kth-order univariate Legendre polynomial is given by $\mathscr{L}_k(x) \equiv \frac{1}{2^k k!}\frac{d^k}{dx^k}\left(x^2-1\right)^k$, and $\mathscr{L}_j(\cdot)$ is orthogonal to $\mathscr{L}_k(\cdot)$ under the Lebesgue measure on $[-1,1]$ for $j \neq k$.[16] To set up the series basis using Legendre polynomials, we first rescale the $X_{1,t}$ (resp. $X_{2,t}$) conditioning variable onto the $[-1,1]$ interval to obtain a transformed variable $\widetilde{X}_{1,t}$ (resp. $\widetilde{X}_{2,t}$). The bivariate series basis is then formed by collecting $\mathscr{L}_j(\widetilde{X}_{1,t})\mathscr{L}_k(\widetilde{X}_{2,t})$ for all $j,k \geq 0$.

It is worth clarifying that our econometric theory does not require the regressors in the series estimation to be orthogonal. The construction above is not meant to achieve orthogonality among regressors, either. Instead, we employ this construction only for the purpose of reducing their empirical correlation so as to improve numerical stability in the practical implementation of series regression (which would be a non-issue if the researcher had infinite numerical precision). To better achieve this goal, a rule-of-thumb is to rescale $X_{j,t}$ in a way such that the empirical distribution of the transformed variable $\widetilde{X}_{j,t}$ is "roughly" uniform on $[-1,1]$, so that we may better exploit the orthogonality property of the Legendre polynomials. Our practical recommendation is to transform $X_{j,t}$ onto $[0,1]$ using its empirical cumulative distribution function, which may be calibrated using any reasonable parameterization (e.g., the normal distribution), and then rescale it linearly onto $[-1,1]$.

Finally, in order to examine how the finite-sample performance of the tests depends on the pre-specified dictionary $\mathcal{M}$, we consider $\mathcal{M} = \{\mathscr{L}_j(x_1)\mathscr{L}_k(x_2) : j,k \geq 0, j+k \leq p\})$ for $p = 2,4,6$, and 8, so that the resulting dictionary contains $m = 6$, 15, 28, and 45 terms, respectively.

[16] The Legendre polynomials can also be computed recursively as $\mathscr{L}_0(x) = 1$, $\mathscr{L}_1(x) = x$, and $\mathscr{L}_k(x) = \frac{2k-1}{k}x\mathscr{L}_{k-1}(x) - \frac{k-1}{k}\mathscr{L}_{k-2}(x)$ for $k \geq 2$.

4.2 Results

We start with discussing the results from the size analysis (i.e., $\delta = 0$). Table 1 presents the finite-sample rejection rates of the selective, non-selective, and uncorrected tests at the 5% significance level under the null hypothesis. Since the results for the $\rho = 0.5$ and 0.8 cases are similar, we shall focus our discussion on the former for brevity.

Panel A of Table 1 shows that the selective test controls size quite well. Specifically, we observe that the test's null rejection rates are generally very close to the 5% nominal level as long as the sample size is not too small (i.e., $n = 250$ or 500), or the dictionary $\mathcal{M}$ is not too large (i.e., $m = 6$ or 15). The only visible size distortion occurs when $\mathcal{M}$ contains "many" series terms and the sample size is small (i.e., $m = 45$ and $n = 150$). That noted, this is actually a quite challenging inferential scenario, because the number of candidate regressors is nearly one third of the sample size. It is perhaps remarkable that the selective test over-rejects only by less than 4% even under this "stress test."

The results for the non-selective test, reported on Panel B, show a sharp contrast. First note that the non-selective test also controls size well for small $\mathcal{M}$ with $m = 6$, which is consistent with the asymptotic theory of Belloni, Chernozhukov, Chetverikov, and Kato (2015) and Li and Liao (2020). However, as m increases to 15, the non-selective test starts to show nontrivial over-rejection (with 23.4% rejection rate) when $n = 150$. We clearly see that this is a small-sample phenomenon, because the size-distortion shrinks quickly as we increase the sample size to $n = 500$. The over-rejection becomes substantially more severe for larger $\mathcal{M}$. Indeed, when $m = 45$, the non-selective test almost always (mistakenly) rejects the null hypothesis when the sample size $n = 150$, and it rejects more than 60% of the time even when $n = 500$.

The size distortion of the non-selective test is perhaps not surprising: Since it always employs all approximating functions in $\mathcal{M}$ for the series estimation, the growing-dimensional asymptotics may not provide an adequate finite-sample approximation when the dimension grows "too fast" relative to the sample size. From this perspective, we see why the selective test may help mitigate this issue, in that the Lasso-assisted selection removes most candidate approximating functions (which are all irrelevant under the null hypothesis), and hence, substantially reduces the "effective dimensionality" of the series inference.

This intuition can be further corroborated by the results shown on Panel C for the uncorrected test. Since the uncorrected test is based on the same Lasso-assisted selection, it also benefits from the aforementioned dimension-reduction effect. Looking at the $m = 45$ case in Panel C, we indeed see that the size distortion of the uncorrected test is much smaller than that of the non-selective

Table 1: Rejection Rates Under the Null Hypothesis

	$\rho = 0.5$			$\rho = 0.8$		
	$n = 150$	$n = 250$	$n = 500$	$n = 150$	$n = 250$	$n = 500$
			Panel A: Selective Test			
$m = 6$	0.053	0.050	0.045	0.054	0.048	0.048
$m = 15$	0.059	0.045	0.043	0.057	0.051	0.048
$m = 28$	0.072	0.055	0.051	0.070	0.055	0.050
$m = 45$	0.088	0.066	0.053	0.078	0.063	0.053
			Panel B: Non-selective Test			
$m = 6$	0.061	0.053	0.050	0.062	0.057	0.046
$m = 15$	0.234	0.143	0.082	0.240	0.144	0.079
$m = 28$	0.712	0.482	0.268	0.720	0.492	0.267
$m = 45$	0.958	0.853	0.620	0.959	0.852	0.619
			Panel C: Uncorrected (Selective) Test			
$m = 6$	0.076	0.067	0.060	0.084	0.073	0.066
$m = 15$	0.143	0.120	0.109	0.148	0.132	0.114
$m = 28$	0.202	0.155	0.136	0.177	0.155	0.141
$m = 45$	0.214	0.177	0.155	0.197	0.169	0.145

Note: This table reports the rejection rates of the selective test, the non-selective test, and the uncorrected selective test at the 5% significance level under the null hypothesis (i.e., $\delta = 0$). These results are generated for a variety of specifications under which the autoregressive coefficient $\rho \in \{0.5, 0.8\}$, the number of candidate basis functions $m \in \{6, 15, 28, 45\}$, and sample size $n \in \{150, 250, 500\}$. The rejection rates are computed based on 10,000 Monte Carlo replications.

test. Nevertheless, the uncorrected test still over-rejects by a nontrivial amount, and hence, is clearly inferior to the proposed selective test in terms of size control. Recall that the selective and the uncorrected tests share the same test statistic $\widehat{T}_{\mathcal{L}}$ and they differ only in the construction of critical values. This comparison thus directly shows the necessity of accounting for the "truncation effect" induced by the data-driven selection.

Overall, the size analysis shows that the proposed selective test has excellent size control, even in adversarial situations with a small sample size and/or a large number of candidate approximating functions. In contrast, the non-selective and the uncorrected tests are able to control size properly only when m is relatively small, and may suffer from severe size distortions in general. The selective test is clearly the most reliable method among the three.

Next, we compare the finite-sample powers of these tests. For brevity, we focus on the setting with $n = 500$. Since the non-selective and uncorrected tests generally suffer from nontrivial size distortions, directly comparing their power with that of the selective test is problematic, as the most size-distorted test may (misleadingly) appear to be the most powerful. We thus instead focus on the size-adjusted power. Figure 1 plots the size-adjusted power curves for the selective, non-selective, and uncorrected tests for $m = 6$ or 45.[17]

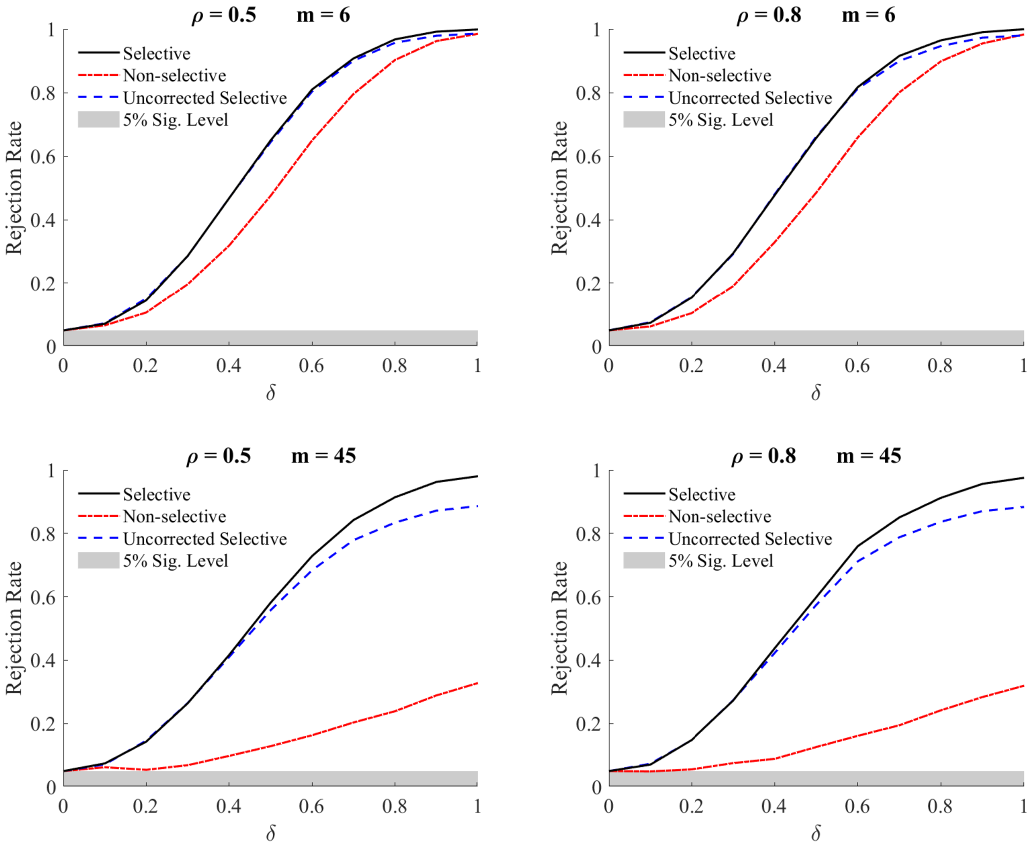
Looking at Figure 1, we first note that the size-adjusted power curves of all three tests hit the nominal size 0.05 at $\gamma = 0$ by construction and, as expected, their rejection rates are increasing in δ; recall that δ quantifies the "distance" between the null and alternative hypotheses. From the top row of the figure, we see that the proposed selective test and the uncorrected test have similar power properties when $m = 6$, and they are more powerful than the benchmark non-selective test. The latter finding is consistent with the intuition that the Lasso-assisted selection helps the tests seek power in a targeted fashion.

The case with "many" series terms (i.e., $m = 45$) displayed on the bottom row of Figure 1 shows a more striking contrast. Indeed, the size-adjusted power of the proposed selective test is far higher than that of the non-selective test, and the former also outperforms the uncorrected test by a notable margin. These findings suggest that the non-selective and uncorrected tests not only suffer from non-trivial size distortions as seen in Table 1, they also deliver worse trade-offs between size and power than the proposed selective test, as measured by the size-adjusted power.

In summary, the simulation study above shows that the proposed selective test has excellent size control across a broad range of scenarios, and is notably more powerful than the non-selective test. These findings clearly demonstrate the usefulness of our proposal relative to that benchmark. We

[17] The cases with $m = 15$ or 28 are bracketed by these two "corner" cases with similar patterns, and so, are omitted for brevity.

Figure 1: Simulation Results: Size-adjusted Power Curves



Note: This figure plots the size-adjusted Monte Carlo rejection rates of the selective test (solid), the non-selective test (dotted), and the uncorrected selective test (dashed) at the 5% significance level (highlighted by the shaded area) over $\delta \in \{0, 0.1, 0.2, \ldots, 1\}$. Results for $m = 6$ (resp. $m = 45$) are reported on the top (resp. bottom) row. Results for $\rho = 0.5$ (resp. $\rho = 0.8$) are reported on the left (resp. right) column. The sample size is fixed at $n = 500$. The rejection rates are computed based on 10,000 Monte Carlo replications.

also see that the "naive" uncorrected selective test generally has nontrivial size distortion, which confirms the necessity of adopting our novel critical value. Given these findings, we unambiguously recommend the selective test for practical applications.[18]

5 Concluding remarks

Conditional moment restrictions may be tested by running a nonparametric series regression. The guidance from the conventional theory is to search for power broadly by using a relatively large number of approximating functions in the series estimation. The cost of doing so could be concerning in practice: If some, even many, regressors are not important for capturing the main features of the conditional expectation function, they may dilute power and, at the same time, distort size. In view of the vast and burgeoning literature on machine-learning-based feature selection, it appears rather natural to use this type of methods, such as Lasso, to select series terms before running the nonparametric test. However, as this paper shows, the data-driven selection itself may cause size distortion through restricting the score on a random polytope (which in turn affects the score's asymptotic normality). This take-home message complements in an interesting way the "orthogonality-induced negligibility" phenomenon articulated by Belloni, Chernozhukov, and Hansen (2014) in a distinct semiparametric context. Our proposed critical value is effective in correcting for this effect. The resulting selective test exhibits improved size and power properties, which is consistent with the theoretical intuition. In this paper, we have focused on the Lasso method for feature selection. The underlying strategy may be applied more broadly to the other variable-selection methods, provided that a tractable characterization of the selection event is available. This seems to be an interesting topic for future research.

Appendix

A.1 Implementation Details

This section provides the additional details related to the implementation of the proposed selective test, which include (i) the exact expressions of $b_{\mathcal{S}}(\mathbf{s})$, $b'_{l,\mathcal{S}}(\mathbf{s})$, $b'_{u,\mathcal{S}}(\mathbf{s})$, $A_{\mathcal{S}}$, and $A'_{\mathcal{S}}$ that are needed in the definition of $\Pi(\mathcal{S}, \mathbf{s}, \lambda_n)$; (ii) a feasible algorithm for determining the Lasso penalty parameters in (2.8).

[18] The finite sample distributions of the number of series terms selected in the Lasso estimation are provided in Section SC.1 of the online supplemental appendix.

Requisite definitions related to the selection event. We provide the precise definitions of $b_{\mathcal{S}}(\mathbf{s})$, $b'_{l,\mathcal{S}}(\mathbf{s})$, $b'_{u,\mathcal{S}}(\mathbf{s})$, $A_{\mathcal{S}}$, and $A'_{\mathcal{S}}$ for a given selection $\mathcal{S}$ satisfying $\mathcal{M}_0 \subseteq \mathcal{S} \subseteq \mathcal{M}$ and a sign configuration $\mathbf{s} \in \{\pm 1\}^{|\mathcal{S}\setminus\mathcal{M}_0|}$. These quantities are used to define the polytope $\Pi(\mathcal{S}, \mathbf{s}, \lambda_n)$ and its proxy $\widehat{\Pi}(\mathcal{S}, \mathbf{s}, \lambda_n)$. Let $\boldsymbol{\omega}_{\mathcal{S}\setminus\mathcal{M}_0}$ and $\boldsymbol{\omega}_{\mathcal{M}\setminus\mathcal{S}}$ denote the subvectors of $\boldsymbol{\omega} \equiv (\omega_j)_{j\in\mathcal{M}_0^c}$ indexed by $\mathcal{S}\setminus\mathcal{M}_0$ and $\mathcal{M}\setminus\mathcal{S}$, respectively. For ease of notation, we write $A^+ \equiv (A^\top A)^{-1}A^\top$ for any matrix A with full column rank and adopt the convention that any matrix indexed by the empty set is empty. The quantities of interest are defined as

$$\begin{cases} b_{\mathcal{S}}(\mathbf{s}) \equiv \operatorname{diag}(\mathbf{s})\,(n^{-1}\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^\top \widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0})^{-1}\operatorname{diag}\left(\boldsymbol{\omega}_{\mathcal{S}\setminus\mathcal{M}_0}\right)\mathbf{s}, \\ b'_{l,\mathcal{S}}(\mathbf{s}) \equiv -\boldsymbol{\omega}_{\mathcal{M}\setminus\mathcal{S}} - \widetilde{\mathbf{P}}_{\mathcal{M}\setminus\mathcal{S}}^\top(\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^+)^\top \operatorname{diag}\left(\boldsymbol{\omega}_{\mathcal{S}\setminus\mathcal{M}_0}\right)\mathbf{s}, \\ b'_{u,\mathcal{S}}(\mathbf{s}) \equiv \boldsymbol{\omega}_{\mathcal{M}\setminus\mathcal{S}} - \widetilde{\mathbf{P}}_{\mathcal{M}\setminus\mathcal{S}}^\top(\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^+)^\top \operatorname{diag}\left(\boldsymbol{\omega}_{\mathcal{S}\setminus\mathcal{M}_0}\right)\mathbf{s}, \\ A_{\mathcal{S}} \equiv \left((n^{-1}\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^\top \widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0})^{-1}, \mathbf{0}_{|\mathcal{S}\setminus\mathcal{M}_0|\times|\mathcal{M}\setminus\mathcal{S}|}\right)\left(-\mathbf{P}_{\mathcal{M}_0^c}^\top(\mathbf{P}_{\mathcal{M}_0}^+)^\top, \mathbf{I}_{|\mathcal{M}_0^c|}\right), \\ A'_{\mathcal{S}} \equiv \left(-\widetilde{\mathbf{P}}_{\mathcal{M}\setminus\mathcal{S}}^\top(\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^+)^\top, \mathbf{I}_{|\mathcal{M}\setminus\mathcal{S}|}\right)\left(-\mathbf{P}_{\mathcal{M}_0^c}^\top(\mathbf{P}_{\mathcal{M}_0}^+)^\top, \mathbf{I}_{|\mathcal{M}_0^c|}\right). \end{cases} \tag{A.1}$$

A data-driven choice of Lasso penalty parameters. We propose a feasible choice of the penalty parameters λ_n and $\{\omega_j\}_{j\in\mathcal{M}_0^c}$ that are needed to implement the Lasso estimation in (2.8). We also show that it satisfies the high-level Assumption 3, and hence, is coherent within our econometric framework. This choice is used in our simulation study, and we recommend it for practical applications. The algorithm is given below, followed by its theoretical justification.

Algorithm A (A Recommended Choice of Penalty Parameters)

Step 1. Run a preliminary Lasso estimation with the resulting coefficient given by

$$\hat{\gamma} \equiv \underset{\gamma\in\mathbb{R}^m}{\operatorname{argmin}} \left\{ \frac{1}{2}\sum_{t=1}^{n}(Y_t - P(X_t)^\top\gamma)^2 + \sqrt{n\log(m)\log(\log(n))}\sum_{j\in\mathcal{M}_0^c}|\gamma_j| \right\}.$$

Step 2. Set the weights in (2.8) as $\omega_j = (|\hat{\gamma}_j| + n^{-1/2})^{-1}$ for each $j \in \mathcal{M}_0^c$.

Step 3. Let k_γ denote the cardinality of $\{j \in \mathcal{M}_0^c : \hat{\gamma}_j \neq 0\}$ and σ_γ^2 denote the sample variance of $Y_t - P\left(X_t\right)^\top \hat{\gamma}$. Set the penalty sequence in (2.8) as $\lambda_n = \sigma_\gamma \max\{k_\gamma^{1/2}, 1\}\log(m)\log(\log(n))$. □

Proposition A1. *Suppose that Assumptions 1 and 2 hold,* $n^{-1}\sum_{t=1}^n \epsilon_t^2 = \sigma_\epsilon^2 + o_p(1)$ *for some positive constant* σ_ϵ^2*, and*

$$\min_{j\in\mathcal{R}} |b_{n,j}^*| \succ |\mathcal{R}|\log\left(n\right)n^{-1/2}. \tag{A.2}$$

Then the penalty parameters λ_n *and* $(\omega_j)_{j\in\mathcal{M}_0^c}$ *described in Algorithm A satisfy Assumption 3 when* $Y_t = Y_t(\theta^*)$*. The same conclusion obtains for* $Y_t = Y_t(\hat{\theta})$*, if Assumption 4 also holds.*

A.2 A pedagogical illustration for the selection effect

In subsection 2.2, we describe how the data-driven selection affects the subsequent inference. The mechanism manifests as the restriction on the score vector $n^{-1/2}\mathbf{P}^\top\boldsymbol{\epsilon}$ within the $\Pi(\mathcal{S},\mathbf{s},\lambda_n)$ polytope, and our proposed critical value adjusts for this truncation effect. That general discussion, however, is somewhat abstract and notationally involved. To further clarify the main force at play, in this subsection we complement the general discussion with a pedagogical example under which the polytope has a simple form.

The setting is as follows. Suppose that X_t is scalar-valued, the dictionary $\mathcal{M}$ contains only two terms, the constant term and the linear term X_t, and the prior choice set $\mathcal{M}_0 = \{1\}$. This means, the constant term is always included in the regression and Lasso determines whether the linear term should be included or not. The specification of the post-selection regression is thus given by

$$\begin{cases} Y_t = a + bX_t + \epsilon_t & \text{if} \quad \mathcal{L} = \mathcal{M}, \\ Y_t = a + \epsilon_t & \text{if} \quad \mathcal{L} = \mathcal{M}_0. \end{cases}$$

The (full) score vector is

$$n^{-1/2}\mathbf{P}^\top\boldsymbol{\epsilon} = \begin{pmatrix} n^{-1/2}\sum_{t=1}^n \epsilon_t \\ n^{-1/2}\sum_{t=1}^n X_t\epsilon_t \end{pmatrix}. \tag{A.3}$$

The inferential analysis for the test concentrates on the null hypothesis, which under the current setting corresponds to $a = b = 0$. Under the null, the collection of $\Pi(\mathcal{S},\mathbf{s},\lambda_n)$ polytopes has the following simple structure: with $\bar{X}$ denoting the sample average $n^{-1}\sum_{t=1}^n X_t$, we have

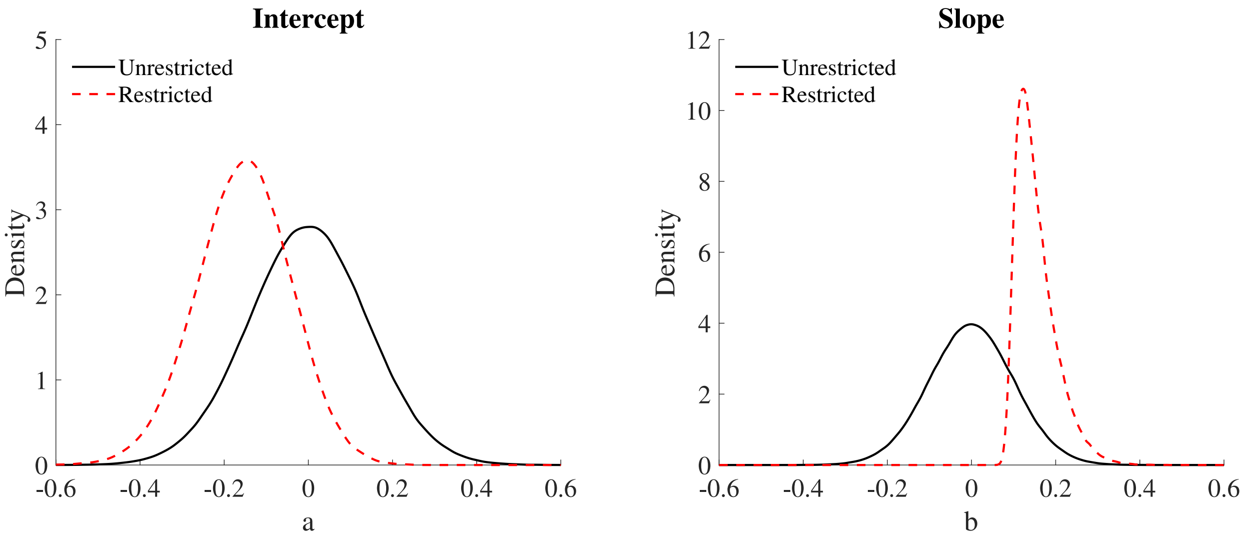
$$\Pi(\mathcal{S},\mathbf{s},\lambda_n) = \begin{cases} \left\{(z_1,z_2) \in \mathbb{R}^2 : z_2 > \bar{X}z_1 + n^{-1/2}\lambda_n\right\}, & \text{if } \mathcal{S} = \mathcal{M},\ \mathbf{s} = +1, \\ \left\{(z_1,z_2) \in \mathbb{R}^2 : z_2 < \bar{X}z_1 - n^{-1/2}\lambda_n\right\}, & \text{if } \mathcal{S} = \mathcal{M},\ \mathbf{s} = -1, \\ \left\{(z_1,z_2) \in \mathbb{R}^2 : \left|z_2 - \bar{X}z_1\right| \leq n^{-1/2}\lambda_n\right\} & \text{if } \mathcal{S} = \mathcal{M}_0. \end{cases} \tag{A.4}$$

Geometrically, the first (resp. second) set corresponds to the half-plane above (resp. below) the line with slope $\bar{X}$ and intercept $n^{-1/2}\lambda_n$ (resp. $-n^{-1/2}\lambda_n$). The third set corresponds to the "stripe" between those two borderlines.

We then readily see how the selection step restricts the score vector. For brevity, we only discuss the case with $\mathcal{S} = \mathcal{M}$ and $\mathbf{s} = +1$; that is, the X_t term is selected by Lasso with a positive sign. Plugging (A.3) and (A.4) into (2.13), we may characterize this selection event explicitly as

$$\left\{n^{-1/2}\sum_{t=1}^n X_t\epsilon_t > \bar{X}\left(n^{-1/2}\sum_{t=1}^n \epsilon_t\right) + n^{-1/2}\lambda_n\right\}, \tag{A.5}$$

Figure 2: Numerical Illustration: Effect of Truncation



Note: This figure plots the distributions of the regression coefficients of $Y_t = a + bX_t + \epsilon_t$. The data is formed as a random sample with sample size $n = 100$ such that $X_t \sim \mathcal{N}(1, 1)$ and $\epsilon_t \sim \mathcal{N}(0, 1)$. The null hypothesis is imposed by setting $a = b = 0$. In the unrestricted case, the least-squares estimation is done without selection. In the restricted case, the distribution is restricted to event on which the X_t term is selected by Lasso with a positive sign. The Lasso penalty parameter is $\lambda = 10$ and the ω weight is normalized to 1.

which shows that the score vector $n^{-1/2}\mathbf{P}^\top \boldsymbol{\epsilon} = (n^{-1/2}\sum_{t=1}^n \epsilon_t, n^{-1/2}\sum_{t=1}^n X_t\epsilon_t)$ is restricted by an inequality constraint once the selection is made. Recall from the "textbook" regression theory that the asymptotic normality of the least-square estimator for (a, b), denoted $(\hat{a}, \hat{b})$, is driven by the asymptotic normality of the score vector $n^{-1/2}\mathbf{P}^\top \boldsymbol{\epsilon}$. But restricting its support via (A.5) clearly will alter its distribution, which in turn will affect that of the least-square estimates. To visualize this effect, we plot in Figure 2 the distributions of the least-square estimates of (a, b) with and without the inequality restriction (A.5) computed in a numerical experiment. From the figure, we see that the selection-induced restriction indeed "pushes" the finite-sample distribution of the least-squares estimates notably away from the benchmark (unrestricted) normal distribution suggested by the "textbook" theory. It is then intuitively clear that this effect will further contaminate the distributions of "functional estimator" $\hat{g}(x) = \hat{a} + \hat{b}x$ and the associated sup-t statistic, and so, leads to size distortion.

The intuition gained from this pedagogical example alludes to a more general logic. Although the population coefficients of all regressors are zero under the null hypothesis, Lasso in any given finite sample may select a few regressors purely due to randomness. Including extra regressors in this manner is far from innocuous, because the regressors are selected precisely because they appear (spuriously) important in the Lasso estimation. As such, they also tend to appear important in the subsequent series estimation, and so, are likely to result in false rejections of the null hypothesis. Put simply, the post-selection test might be systematically "fooled by randomness" in finite samples due to this mechanism, whenever the theoretically optimistic "oracle" property is not actually in force. It is worth clarifying that this concern does not apply to the familiar scenario in which an empiricist a priori decides to include a few exogeneous control variables in the regression, say, based on economic reasoning rather than statistical screening. The key lesson here is that the selection step may make the originally exogeneous variables "effectively endogenous" by truncating the support of the score. The conditional adjustment underlying our proposed critical value is exactly designed to account for this truncation effect. This adjustment indeed provides excellent finite-sample performance as we have shown in the more comprehensive numerical experiment presented in Section 4.

References

Andrews, D. W. K. (1991a): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, 59(2), 307–345.

——— (1991b): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817–858.

——— (1994): "Chapter 37 Empirical Process Methods in Econometrics," vol. 4 of *Handbook of Econometrics*, pp. 2247–2294. Elsevier, Amsterdam, Netherlands.

Andrews, D. W. K., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81(2), 609–666.

Andrews, I., T. Kitagawa, and A. McCloskey (2021a): "Inference After Estimation of Breaks," *Journal of Econometrics*, 224(1), 39–59.

——— (2021b): "Inference on Winners," Discussion paper, Harvard University, UCL, and University of Colorado.

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

Belloni, A., and V. Chernozhukov (2011): "High Dimensional Sparse Econometric Models: An Introduction," in *Inverse Problems and High-Dimensional Estimation*, pp. 121–156. Springer.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186(2), 345 – 366, High Dimensional Problems in Econometrics.

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81(2), 608–650.

BICKEL, P. J., Y. RITOV, AND A. B. TSYBAKOV (2009): "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37(4), 1705 – 1732.

BOTEV, Z. I. (2016): "The Normal Law Under Linear Restrictions: Simulation and Estimation via Minimax Tilting," *arXiv preprint arXiv:1603.04166*.

BROWN, B. W., AND S. MAITAL (1981): "What do Economists Know? An Empirical Study of Experts' Expectations," *Econometrica*, 49(2), 491–504.

CHEN, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6B, chap. 76. Elsevier, 1 edn.

CHEN, X., AND T. M. CHRISTENSEN (2015): "Optimal Uniform Convergence Rates and Asymptotic Normality for Series Estimators Under Weak Dependence and Weak Conditions," *Journal of Econometrics*, 188(2), 447–465.

CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

CHODOROW-REICH, G., AND L. KARABARBOUNIS (2016): "The Cyclicality of the Opportunity Cost of Employment," *Journal of Political Economy*, 124(6), 1563–1618.

COX, G., AND X. SHI (2019): "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models," *arXiv preprint arXiv:1907.06317*.

HANSEN, L. P., AND R. J. HODRICK (1980): "Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis," *Journal of Political Economy*, 88(5), 829–853.

HANSEN, L. P., AND K. J. SINGLETON (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50(5), 1269–1286.

LEE, J. D., D. L. SUN, Y. SUN, AND J. E. TAYLOR (2016): "Exact Post-Selection Inference, With Application to the Lasso," *The Annals of Statistics*, 44(3), 907–927.

LI, J., AND Z. LIAO (2020): "Uniform Nonparametric Inference for Time Series," *Journal of Econometrics*, 219(1), 28–51.

LI, J., Z. LIAO, AND R. QUAEDVLIEG (2020): "Conditional Superior Predictive Ability," *Review of Economic Studies, forthcoming.*

LIAO, Z., AND X. SHI (2020): "A Nondegenerate Vuong Test and Post Selection Confidence Intervals for Semi/nonparametric Models," *Quantitative Economics*, 11(3), 983–1017.

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

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics, IV, Edited by R. F. Engle and D. L. McFadden*, pp. 2112–2245.

Pollard, D. (2001): *A User's Guide to Measure Theoretic Probability.* Cambridge University Press.

Romer, C. D., and D. H. Romer (2000): "Federal Reserve Information and the Behavior of Interest Rates," *American Economic Review*, 90(3), 429–457.

Stock, J. H., and J. H. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68(5), 1055–1096.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

van der Vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Processes.* Springer-Verlag.

Yuan, M., and Y. Lin (2006): "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.

Zhao, P., and B. Yu (2006): "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

Zou, H. (2006): "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101(476), 1418–1429.

Zou, H., and T. Hastie (2005): "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

Zou, H., and H. H. Zhang (2009): "On the Adaptive Elastic-Net With a Diverging Number of Parameters," *Annals of statistics*, 37(4), 1733.