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

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

June 30, 2021

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, Frank Schorfheide and seminar participants at 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

Conditional moment restrictions in the form of $\mathbb{E}\left[Y_t|X_t\right] = 0$ are of central importance in econometric analysis. These restrictions may arise from Euler or Bellman equations in dynamic equilibrium models (Hansen (1982), Hansen and Singleton (1982)), independence-type conditions for instrumental variables (Amemiya (1977), Chamberlain (1987), Newey (1990)), rational expectations (Muth (1961), Hansen and Hodrick (1980), Brown and Maital (1981)), or the equal conditional performance of competing economic forecasts (Giacomini and White (2006)). In structural problems, the "residual" Y_t may depend on an unknown finite-dimensional parameter, such as preference parameters in a utility function. In this context, a large literature has been devoted to developing the semiparametrically efficient estimation for the parameter by exploiting the full information content embedded in the conditional moment restrictions (Newey (1990), Ai and Chen (2003), Donald, Imbens, and Newey (2003), Kitamura, Tripathi, and Ahn (2004)). The efficient estimation evidently relies on the validity of the underlying restrictions, and hence, naturally motivates testing $\mathbb{E}\left[Y_t|X_t\right] = 0$ as a specification check (Bierens (1982, 1990), Hardle and Mammen (1993), Hong and White (1995), Fan and Li (1996), Horowitz and Spokoiny (2001), Donald, Imbens, and Newey (2003)). More generally, even in the absence of unknown parameters, the testing problem is still of great empirical relevance for studying rational-expectation or equal-predictive-ability hypotheses in the context of forecast evaluation. This paper proposes a new test for conditional moment restrictions in a general setting with heterogeneous serially dependent data that readily accommodates the aforementioned applications in cross-sectional, time-series, and panel settings.[1] The key novelty of our proposal is to extract the main features of the conditional expectation function via a machine-learning-style variable selection (for power improvement) and properly design the critical value to account for the effect of data-driven selection (for size control).

The hypothesis of interest is generally nonparametric in nature as it pertains to the global property of the conditional expectation function $g(\cdot) \equiv \mathbb{E}\left[Y_t|X_t = \cdot\right]$. Instead of attacking the nonparametric problem directly, empiricists often take a shortcut by running a linear regression of Y_t on X_t and test whether the coefficients are zero.[2] The drawback of this parametric approach

[1] The asymptotic theory of this paper is applicable for general heterogeneous mixingales, which includes martingale differences, linear processes, various mixing concepts, and certain near-epoch dependence as special cases; see Davidson (1994) for a comprehensive review on mixingales and various dependence concepts. It is well known that mixingales form a far more general class of processes than those characterized by various mixing concepts. For example, Andrews (1984, 1985) showed constructively that even nearly independent triangular arrays are not strongly mixing; as a result of the ranking of mixing coefficients, they are not ρ-mixing or β-mixing, either. Our results are not subject to Andrews's critique.

[2] As a case in point, Romer and Romer (2000) tests the rationality of the Federal Reserve's Greenbook forecasts

is that it cannot reveal nonlinear violations of the null hypothesis. A natural way to address this issue is to include additional nonlinear transformations of X_t, say polynomials or splines, into the regression (Eubank and Spiegelman (1990)). In the thought experiment that the number of approximating terms grows with the sample size, the latter approach may be given a nonparametric interpretation by invoking the series estimation theory (Andrews (1991a), Newey (1997), Chen (2007)). The nonparametric test may then be carried out by checking whether the functional estimator for $g(\cdot)$ is statistically zero in a uniform sense.[3] In addition, since the series estimation is effectively a least-squares regression with a growing number of regressors, the test can also be used to examine whether Y_t is linearly dependent on a "large" number of conditioning variables (i.e., the dimensionality of X_t grows asymptotically). The econometric method proposed in this paper is equally applicable in both nonparametric and many-regressor settings, but we shall mostly focus on the former for ease of discussion.

The series-based test could be "tricky" to apply because practitioners are often unsure about which approximating functions should be included in the series regression. Indeed, econometric theory often does not provide direct guidance on this choice but only specifies how fast the number of series terms needs to grow with the sample size.[4] Nevertheless, it is intuitively clear what might go wrong when the series terms are chosen improperly. The risk of using "too few" approximating terms is evident, that is, the series estimator may miss the main features of $g(\cdot)$, and hence, has little power against certain alternatives. The consequence of including "too many" series terms might be ambiguous. First note that, while a highly flexible specification may give the test nontrivial power in a broad range of "directions," its power may be low in any specific direction. For example, if the conditional expectation function happens to have a simple form $g(x) = 1+x$ under the alternative, including (a large number of) higher-order polynomials can only add statistical noise and so dilutes the power of the test. One might thus expect the series-based test to have omniscient but low

by regressing the forecasting error on the forecast itself; also see Hansen and Hodrick (1980) and Brown and Maital (1981) for applications in the other contexts.

[3] Uniform inference under the growing-dimensional setting is somewhat nonstandard because it is a non-Donsker problem. This technical issue may be addressed by using the growing-dimensional strong approximation, or coupling, theory recently developed by Belloni, Chernozhukov, Chetverikov, and Kato (2015) and Li and Liao (2020) for serially independent and dependent data, respectively.

[4] An important exception is put forth by Donald, Imbens, and Newey (2009) (also see Donald and Newey (2001)), who propose an optimal choice of the number of instruments to minimize the mean-squared-error in the estimation of the unknown finite-dimensional parameter identified by the conditional moment restrictions. But Donald et al. also note that their selection method does not generally provide the most accurate inference. Also note that our econometric interest is not on the unknown parameter per se, but rather on the nonparametric testing of the conditional moment restrictions. The method of Donald, Imbens, and Newey (2009) is thus not applicable in our present context.

power when a large number of series terms are used. In finite samples, however, this power-dilution effect may be counteracted by a size-distortion effect, which stems from the fact that, when there are many regressors, the asymptotic Gaussian approximation (possibly in the form of a growing-dimensional Gaussian coupling) may behave poorly when the sample size is moderate and so cause over-rejection. When this effect is in force, tests implemented with a large number of series terms could appear "very powerful" but unfortunately for a very wrong reason.

These heuristics also suggest that the (counteracting) power-dilution and size-distortion effects may actually be mitigated at the same time if we can select a relatively small set of series terms to capture the "main features" of the a priori unknown $g(\cdot)$ function. If so, the test will seek power in a targeted fashion, and the reduced dimensionality also ensures a better finite-sample performance of the growing-dimensional Gaussian approximation. Since the series regression is a least-squares procedure, it is natural for our purpose to employ the Lasso method (Tibshirani (1996)) for selecting series terms.

In a different context, Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2014) have also used the Lasso to select regressors for fitting the conditional expectation functions of the outcome variable and/or endogenous variables given a set of instruments. While we focus on nonparametrically testing whether a conditional expectation function is zero, these authors are interested in a very different question, namely the semiparametrically efficient estimation of a linear treatment effect, for which the conditional expectation function is merely a nonparametric nuisance. As Belloni et al. eloquently explain, their analysis exploits an important orthogonality property of the (semiparametric) problem, so that moderate model selection mistakes do not prevent valid inference about low-dimensional structural parameters. In their context, the sampling uncertainty of the data-driven selection is negligible.

However, there is no reason to expect that the "negligibility" property would obtain in our current setting. In fact, the notion of "orthogonality" is irrelevant for our analysis because the conditional expectation function is exactly our inferential target, rather than a nuisance. This concern turns out to be highly relevant: Our simulation results show that tests based on Lasso-selected series terms exhibit quite nontrivial size distortion (e.g., 15%–20% finite-sample rejection rates for 5%-level tests). The theoretical heuristic and the corroborating Monte Carlo evidence suggest that the sampling variability in the data-driven selection may have a nontrivial impact on the series-based nonparametric test.

Set against this background, a key component of our theoretical analysis is to understand how the "first-stage" feature selection affects the "second-stage" testing, and then propose a corrected critical value to account for it. We rely on an analytical characterization of the selection event (i.e., the event when a specific subset of regressors are selected by Lasso) as a system of linear-inequality restrictions on the score vector of the series regression. This characterization reveals that the score is restricted on a random polytope once the data-driven selection is made.[5] Since the asymptotics of the series estimator is captured by the (growing-dimensional) Gaussian coupling for the score, restricting its support on a polytope effectively results in a form of truncated normality. This explains why treating the data-driven selection "as given" and deriving critical values from the conventional Gaussian approximation may lead to size distortion as seen in the simulations. Constructively, this characterization also allows us to design a new critical value that accounts for the truncation effect. In the simulations, we find that the adjusted critical value greatly improves the test's size control in finite samples. To our knowledge, our method provides the first attempt to account for the distorting effect of data-driven feature selection for testing conditional moment restrictions (while also allowing for general serial dependence in the data).

Under some standard regularity conditions, we show theoretically that the proposed selective test controls size asymptotically under the null hypothesis. We also characterize local alternatives against which the test is consistent. An interesting aspect of our power analysis is that it 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 sense that $g(\cdot)$ can be well approximated by a smaller set of series terms. In the extreme case when $g(\cdot)$ 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. Our power analysis is not restricted to this type of "sparse" alternatives. If $g(\cdot)$ loads on a growing number of approximating terms, the theory also shows how the test's rate of consistency drops as $g(\cdot)$ becomes more complex. In the worst-case scenario in which $g(\cdot)$ is "very complex," 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.

This paper is related to several strands of literature in econometrics and statistics. The most closely related is the classical work on series-based nonparametric specification testing, which was first introduced by Eubank and Spiegelman (1990) in a Gaussian model, and then substantially generalized by Hong and White (1995) and Donald, Imbens, and Newey (2003) for non-Gaussian i.i.d. data. In a general time-series setting with heterogeneous mixingales, Li and Liao (2020) propose a test using a uniform inference method justified by a growing-dimensional coupling theory, which we adopt here as well. Compared to the prior work, the key innovation of the present paper is to introduce a machine-learning-style feature selection before running the nonparametric test and,

[5] A polytope is defined as an intersection of half-spaces.

perhaps more importantly, we analyze how the data-driven selection may affect the subsequent test and propose a novel critical value to account for it. Our proposal is thus very different from the existing methods both operationally and theoretically.

Instead of series approximation, kernel smoothing may also be used to carry out the nonparametric test; see, for example, Hardle and Mammen (1993), Fan and Li (1996), and Horowitz and Spokoiny (2001). Our focus is conceptually tangential to the kernel approach because, after all, our analysis starts with the Lasso-based feature selection (instead of, say, a data-driven selection of bandwidth). An advantage of using the series approach is that it allows us to build on the insight and techniques from the vast literature on Lasso and, more generally, growing-dimensional statistics. This setup also readily accommodates the many-regressor setting under which the selection would inform the choice among a range of conditioning variables; this is relevant in empirical work but does not appear to have a direct analogue under the kernel approach.

The series and kernel methods both involve nonparametrically estimating the unknown conditional expectation function. Yet another testing strategy has been developed by Bierens (1982), Bierens (1990), Bierens and Ploberger (1997), among others, which is based on transforming the conditional moments into a continuum of unconditional ones using a set of weight functions (i.e., instruments). The asymptotic properties of the test are then analyzed using a standard empirical-process theory. Bierens's approach is largely complementary to the aforementioned nonparametric-estimation-based methods; see Bierens and Ploberger (1997), Fan and Li (2000), and Horowitz and Spokoiny (2001) for their relative merits. It is worth noting that the asymptotic theory underlying Bierens's tests relies crucially on the tightness of the empirical process associated with the instrumented unconditional moments (i.e., the Donsker property holds), and hence, cannot directly accommodate the setting with a growing number of conditioning variables. Moreover, to the best of our knowledge, the existing literature also appears to be silent on how one might choose the set of weight functions for Bierens-type tests.[6] Donald, Imbens, and Newey (2003) point out an important connection between Bierens's approach and the series approach, that is, the growing number of series terms may also be used as weight functions to form the unconditional moments. From this perspective, our data-driven selection and the accompanying non-Donsker inference technique may be used to extend Bierens's approach. This might be an interesting topic for future research, but is beyond the scope of the present paper.

[6] For example, the theory of Bierens and Ploberger (1997) only requires the set of nuisance parameters indexing the weight functions to have *positive* Lebesgue measure. But it is quite clear that the choice of this index set matters a great deal for the finite-sample performance of the test. Indeed, if the set is very small (and so the weight functions are very similar), the test is effectively based on a single instrument, and hence, may lack power against general alternatives. Having a large index set is analogous to using a large number of series terms.

This paper is also related to the large literature on Lasso (Tibshirani (1996)). The vast majority of work has concentrated on point-estimation properties such as consistency for prediction and variable selection (see, e.g., Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Zou (2006), van de Geer (2008), Zhang and Huang (2008), Bickel, Ritov, and Tsybakov (2009)), but less work has been done for econometric inference. Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2014) have made important contributions in this direction by studying the efficient estimation of a linear treatment effect in certain semiparametric (cross-sectional) contexts.[7] As emphasized above, our nonparametric testing problem is fundamentally different from the semiparametric efficient estimation problem studied by Belloni et al., particularly due to the fact that the conditional expectation function is our primary inferential target, but it is only a nuisance in the semiparametric context. Related to this difference, we also find that the effect of data-driven selection is nontrivial and it is important to properly account for the resulting truncation effect so as to maintain good size control. Our "non-negligibility" finding complements the "orthogonality-induced negligibility" message articulated by Belloni, Chernozhukov, and Hansen (2014). It is also worth noting that most prior work on Lasso is done for the cross-sectional setting (sometimes also under Gaussian and/or homoskedasticity assumptions), whereas our theory is developed for general heterogeneous serially dependent data so as to accommodate a broad range of applications in economics.[8]

Finally, we note that our study may be related to the literature on conditional-moment-based semiparametric inference; see, for example, Kelejian (1971), Amemiya (1977), Hansen (1982), Hansen and Singleton (1982), Newey (1990), Ai and Chen (2003), Donald, Imbens, and Newey (2003), Kitamura, Tripathi, and Ahn (2004), and Ackerberg, Chen, Hahn, and Liao (2014). Our proposed test is complementary to this literature in two ways. Firstly, our test may be used to check the validity of conditional moment restrictions post-estimation. Secondly, when the parameter in the structural model is weakly identified, one may invert the proposed nonparametric specification test for the conditional moment restrictions to obtain an Anderson–Robin confidence set for the unknown parameter. This is analogous to Stock and Wright's (2000) method of inverting the

[7] The analysis of Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2014) may be viewed as conducting inference on a low-dimensional parameter in a high-dimensional model, and hence, can be further related to Van de Geer, Bühlmann, Ritov, Dezeure, et al. (2014), Zhang and Zhang (2014), Javanmard and Montanari (2014a,b), and Lee, Sun, Sun, and Taylor (2016).

[8] Bai and Ng (2008) provide one of the first applications of Lasso in time-series econometrics. Their analysis is mainly empirical, with the main goal of selecting relevant predictors for factor-based forecasting models. In contrast, we are interested in the very different problem of developing a formal testing procedure for conditional moment restrictions. Our test may be applied in the forecast evaluation context for testing forecast rationality (Hansen and Hodrick (1980)) or conditional equal predictive ability (Giacomini and White (2006)).

over-identification test for a given set of unconditional moments.

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 of the proposed method. Technical proofs 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 $\mathcal{X} \subseteq \mathbb{R}^d$.[9] 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 a few 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,

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

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 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 (Conditional Equal Predictive Ability). Giacomini and White (2006) study a conditional forecast comparison problem. Suppose that there are two competing forecasting methods $F_{t+1|t}^{(1)}$ and $F_{t+1|t}^{(2)}$ for $F_{t+1}^{\dagger}$. Given a loss function $L(\cdot,\cdot)$, the relative performance of these forecasts may be quantified by the loss differential $Y_t = L(F_{t+1|t}^{(1)}, F_{t+1}^{\dagger}) - L(F_{t+1|t}^{(2)}, F_{t+1}^{\dagger})$. The null hypothesis in (2.1) suggests that the two forecasting methods have equal expected performance across all conditioning states.

Example 3 (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 $U'(\cdot,\gamma)$ 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 $Y_t(\theta^*) = \frac{\beta U'(C_{t+1},\gamma)}{U'(C_t,\gamma)}R_{t+1} - 1$. Similar equilibrium conditions can also be derived from Bellman equations; see, for example, Li and Liao's (2020) empirical analysis on the search-and-matching model for unemployment (Mortensen and Pissarides (1994)). In the macroeconomic setting, the parameter θ^* is often, though not always, calibrated based on external data and auxiliary models (see, e.g., Chodorow-Reich and Karabarbounis (2016)).

Example 4 (Instrumental Variable Problems) Conditional model restrictions have also been extensively studied in cross-sectional and panel data settings (see, e.g., Amemiya (1977), Chamberlain (1987), Newey (1990), and Donald, Imbens, and Newey (2003, 2009)). The econometric model has the form $\mathbb{E}\left[\rho\left(Z_t,\theta^*\right)|X_t\right] = 0$, where $\rho\left(Z_t,\theta^*\right)$ is the residual of an estimation equation and X_t is an instrument possibly motivated by some natural experiments (see, e.g., Angrist and Krueger (1991)). The unknown parameter θ^* may be semiparametrically efficiently estimated at the $n^{1/2}$ rate of convergence. With $Y_t(\theta^*) = \rho\left(Z_t,\theta^*\right)$, testing (2.1) amounts to nonparametrically testing the validity of the instrumental variables.

As seen from these examples, the testing problem (2.1) may be concerned with the rationality and relative performance of economic forecasts, the specification of a dynamic equilibrium model, or the validity of instrumental variables. 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 $g(\cdot)$. 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 $\mathbb{E}[Y_t] = 0$. 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.[10] 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 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.

[10] 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.

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. As alluded in the Introduction, this is not surprising (at least ex post) because the "orthogonality-induced negligibility" insight of Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2014) does not apply here. 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.[11] 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 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),

[11] 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).

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 the Appendix, 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).[12] 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

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

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.

[12] 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 $\mathcal{S}$.

More notation is needed. Let $\mathbf{I}_n$ denote the n-dimensional identity matrix, $\boldsymbol{\epsilon} \equiv (\epsilon_t)_{1\leq t\leq n}$, and $\mathbf{G} \equiv (g(X_t))_{1\leq t\leq n}$. 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.[13] 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

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

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

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

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 the Appendix 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

[13] 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.

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 $\widehat{Q}_{\mathcal{S}}^{-1}\widehat{\Sigma}_{\mathcal{S}}\widehat{Q}_{\mathcal{S}}^{-1}$.[14] The n-dimensional vector $\mathbf{G}$ 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

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

(2.14), that is,

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

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

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

[15] 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).

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.[16] 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.

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*

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

(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.,

[16] 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 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.

$m \to \infty$), 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.[17] 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 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

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

[17] 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.

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.[18]

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.

Assumption 3 mainly ensures that the Lasso estimator described in (2.8) is sign-consistent under the null and alternative hypotheses.[19] This condition is high-level in nature and it does not directly pin down any specific penalty scheme. In the Appendix, 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 $a_n \ge c_n b_n$ for some strictly positive sequence $c_n\to\infty$.

[18] 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.

[19] 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, $\sqrt{n\log(m)}/(\lambda_n\min_{j\in\mathcal{M}_0^c}\omega_j) = o_p(1)$.

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 in which $g_n(\cdot)$ under the alternative has a parametric but a priori unknown form. In this case, $\kappa\left(\mathcal{M}^\star\right)$ is bounded and the $\kappa\left(\mathcal{M}^\star\right) \log(n)^{1/2} n^{-1/2}$ 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

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

Finally, we set

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

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.[20]

[20] 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.

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 $h_n(x) \equiv \mathbb{E}\left[\partial_\theta Y_t(\theta^*) \mid X_t = x\right]$ 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.[21] 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 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.

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

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 the Appendix, 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 $\mathscr{L}_k\left(x\right) \equiv \frac{1}{2^k k!}\frac{d^k}{dx^k}\left(x^2-1\right)^k$, and $\mathscr{L}_j\left(\cdot\right)$ is orthogonal to $\mathscr{L}_k\left(\cdot\right)$ under the Lebesgue measure on $[-1,1]$ for $j \neq k$.[22] 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.

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

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

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.

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 $n = 500$.

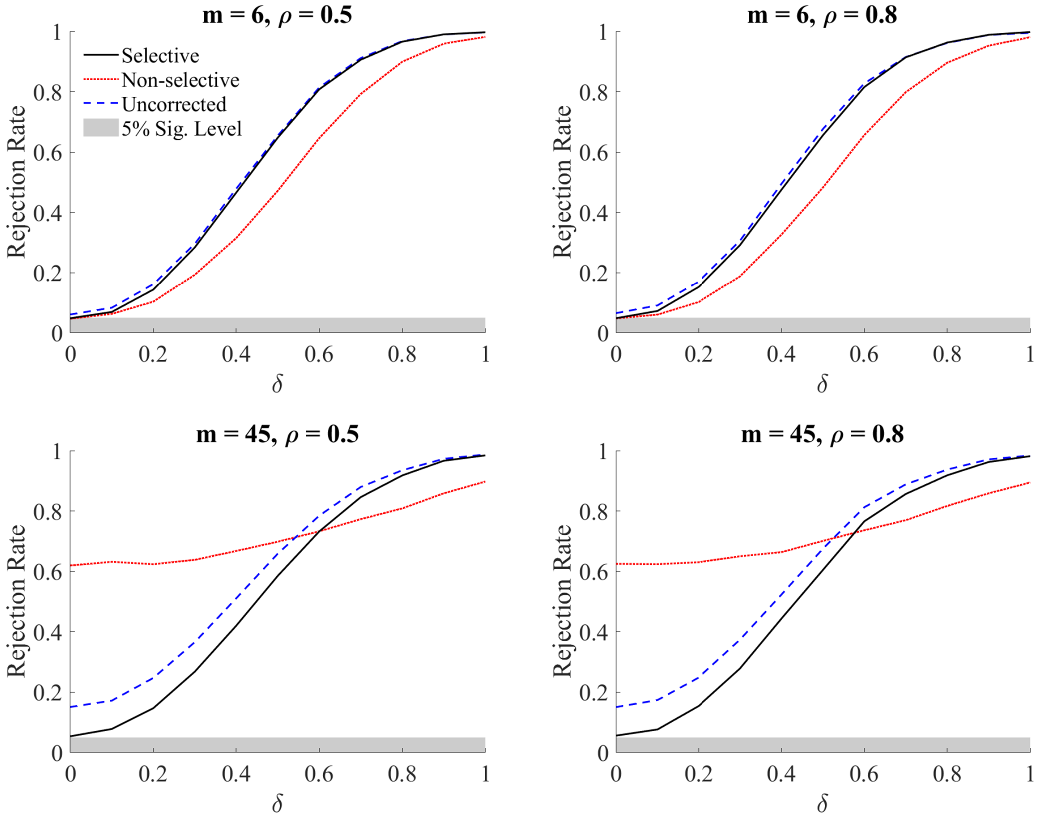
The size distortion of the non-selective test is perhaps not surprising: Since it always employs all approximating functions in $\mathcal{M}$ for the series estimation, the growing-dimensional asymptotics may not provide an adequate finite-sample approximation when the dimension grows “too fast” relative 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 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.

We next turn to the tests’ finite-sample powers. In view of the non-selective test’s rather severe size distortion in small samples, we only focus on the $n = 500$ “large-sample” case so as to have a more meaningful power comparison. This precludes the uninteresting scenario in which

Figure 1: Simulation Results: Power Curves



Note: This figure plots the 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.

the most size-distorted test may (misleadingly) appear to be the most powerful. In Figure 1, we plot the rejection rates of the selective, non-selective, and uncorrected tests as functions of the δ parameter. Recall that δ quantifies the "distance" between the alternative and the null hypotheses, and hence, we expect the rejection rates to be increasing in δ. The left and right columns of the figure correspond to the $\rho = 0.5$ and 0.8 cases, respectively. Regarding the size of the dictionary $\mathcal{M}$, we focus on the two corner cases, $m = 6$ and 45, which are displayed on the top and bottom rows of the figure.[23]

We first investigate the $m = 6$ case plotted on the top panel of Figure 1. In this case, all three tests control size reasonably well, as the left ends of the power curves are all "tied" to the 5% nominal level. It is thus straightforward to interpret the power comparison. Specifically, we see that the selective and the uncorrected tests have nearly identical power profiles and they are notably more powerful than the non-selective test. This is consistent with the intuition that the Lasso-assisted selection helps the test seek power in a targeted fashion.

Results for the $m = 45$ case, which are displayed on the bottom panel of Figure 1, need to be interpreted more carefully, because the non-selective and the uncorrected tests, especially the former, are quite over-sized under the null. Indeed, we see clearly from the figure that the left ends of their power curves are far above the 5% nominal level, whereas the power curve of the selective test "starts" correctly at (almost) 5%. As δ increases, all three tests reject more often. The power curve of the uncorrected test lies above that of the proposed selective test. But, evidently, the former's seemingly better power needs to be taken with a grain of salt, as this is mainly due to its size distortion. Consistent with this logic, we see that the gap between the two power curves shrinks as δ increases, suggesting that the proposed selective test gradually overcomes the "unfair disadvantage" stemming from the uncorrected test's over-rejection under the null. This catch-up effect is even more evident in the comparison between the selective test and the non-selective test: In this case, the rejection rate of the former eventually exceeds that of the latter (as their power curves cross), despite the fact that the non-selective test is rather severely over-sized under the null. This finding thus provides a fairly strong piece of evidence for the efficiency gain from using the proposed selective test, as predicted by our econometric theory.

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. A particularly remarkable finding is that the selective test not only controls size much better than the non-selective test in general, but it can also be more powerful provided that the alternative

[23] As in the size analysis (see Table 1), results for the $m = 15$ and 28 cases are "bracketed" by those from the two corner cases. Those results do not shed additional light on our discussion and so are omitted for brevity.

is sufficiently distant from the null (so as to overshadow the non-selective test's size distortion). These findings clearly demonstrate the usefulness of our proposal relative to that benchmark. We 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.

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: Implementation Details

This appendix 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).

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

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

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*

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

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.

References

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014): "Asymptotic Efficiency of Semiparametric Two-step GMM," *The Review of Economic Studies*, 81(3), 919–943.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Amemiya, T. (1977): "The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model," *Econometrica*, 45(4), 955–968.

Andrews, D. W. K. (1984): "Non-Strong Mixing Autoregressive Processes," *Journal of Applied Probability*, pp. 930–934.

——— (1985): "A Nearly Independent, but Non-Strong Mixing, Triangular Array," *Journal of Applied Probability*, pp. 729–731.

——— (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.

Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

Bai, J., and S. Ng (2008): "Forecasting Economic Time Series Using Targeted Predictors," *Journal of Econometrics*, 146(2), 304–317.

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.

Bierens, H. J. (1982): "Consistent Model Specification Tests," *Journal of Econometrics*, 20, 105–134.

Bierens, H. J. (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58(6), 1443–1458.

Bierens, H. J., and W. Ploberger (1997): "Asymptotic Theory of Integrated Conditional Moment Tests," *Econometrica*, 65(5), pp. 1129–1151.

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.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34(3), 305–334.

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.

Davidson, J. (1994): *Stochastic Limit Theory*. Oxford University Press.

Donald, S. G., G. W. Imbens, and W. K. Newey (2003): "Empirical Likelihood Estimation and Consistent Tests With Conditional Moment Restrictions," *Journal of Econometrics*, 117(1), 55–93.

Donald, S. G., G. W. Imbens, and W. K. Newey (2009): "Choosing Instrumental Variables in Conditional Moment Restriction Models," *Journal of Econometrics*, 152(1), 28–36.

DONALD, S. G., AND W. K. NEWEY (2001): "Choosing the Number of Instruments," *Econometrica*, 69(5), 1161–1191.

EUBANK, R. L., AND C. H. SPIEGELMAN (1990): "Testing the Goodness of Fit of a Linear Model Via Nonparametric Regression Techniques," *Journal of the American Statistical Association*, 85(410), 387–392.

FAN, Y., AND Q. LI (1996): "Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms," *Econometrica*, 64(4), 865–890.

——— (2000): "Consistent Model Specification Tests: Kernel-Based Tests versus Bierens' ICM Tests," *Econometric Theory*, 16(6), 1016–1041.

GIACOMINI, R., AND H. WHITE (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74(6), 1545–1578.

HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

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.

HARDLE, W., AND E. MAMMEN (1993): "Comparing Nonparametric Versus Parametric Regression Fits," *The Annals of Statistics*, 21(4), 1926–1947.

HONG, Y., AND H. WHITE (1995): "Consistent Specification Testing Via Nonparametric Series Regression," *Econometrica*, 63(5), 1133–1159.

HOROWITZ, J. L., AND V. G. SPOKOINY (2001): "An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model against a Nonparametric Alternative," *Econometrica*, 69(3), 599–631.

JAVANMARD, A., AND A. MONTANARI (2014a): "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *The Journal of Machine Learning Research*, 15(1), 2869–2909.

——— (2014b): "Hypothesis Testing in High-Dimensional Regression Under the Gaussian Random Design Model: Asymptotic Theory," *IEEE Transactions on Information Theory*, 60(10), 6522–6554.

KELEJIAN, H. H. (1971): "Two-Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in the Endogenous Variables," *Journal of the American Statistical Association*, 66(334), 373–374.

KITAMURA, Y., G. TRIPATHI, AND H. AHN (2004): "Empirical Likelihood-Based Inference in Conditional Moment Restriction Models," *Econometrica*, 72(6), 1667–1714.

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.*

Meinshausen, N., and P. Bühlmann (2006): "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34(3), 1436 – 1462.

Mortensen, D. T., and C. A. Pissarides (1994): "Job Creation and Job Destruction in the Theory of Unemployment," *The Review of Economic Studies*, 61(3), 397–415.

Muth, J. F. (1961): "Rational Expectations and the Theory of Price Movements," *Econometrica*, 29(3), 315–335.

Newey, W. K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58(4), 809–837.

——— (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 de Geer, S., P. Bühlmann, Y. Ritov, R. Dezeure, et al. (2014): "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *Annals of Statistics*, 42(3), 1166–1202.

van de Geer, S. A. (2008): "High-Dimensional Generalized Linear Models and the Lasso," *The Annals of Statistics*, 36(2), 614 – 645.

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.

Zhang, C.-H., and J. Huang (2008): “The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression,” *The Annals of Statistics*, 36(4), 1567 – 1594.

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

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.