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

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

October 11, 2024

[*] We thank a co-editor and three anonymous referees for their comments which have greatly improved the paper. We also thank Francis Diebold, Xu Cheng, Gregory Cox, Bruce Hansen, Ming Li, Jack Porter, Frank Schorfheide, Xiaoxia Shi, Kenneth West, Zhenlin Yang, Jun Yu and seminar participants at the Chinese University of Hong Kong at Shenzhen, Michigan State University, Tsinghua University, University of Chinese Academy of Sciences, Singapore Management University, University of California-Irvine, University of Warwick, University of Wisconsin-Madison, University of Manchester, University of Pennsylvania for their comments.

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

[‡] Department of Economics, UCLA, Los 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.

Abstract

We develop a new test for conditional moment restrictions via nonparametric series regression, with approximating functions selected by Lasso. A key novelty of our approach is to account for the effect of the data-driven selection, yielding a new critical value constructed on the basis of a nonstandard truncated-Gaussian asymptotic approximation. We show that the test is correctly sized and attains a well-defined sense of adaptiveness that generally results in better power than existing methods. The improvement afforded by the new test is demonstrated in a Monte Carlo study and an empirical application on the conditional evaluation of inflation forecasts.

Keywords: conditional moments; Lasso; machine learning; series estimation; uniform inference; variable selection.

JEL Codes: C14, C22.

1 Introduction

Testing conditional moment restrictions is an important topic in econometrics. One approach is to nonparametrically estimate the conditional moment function via a series regression ([1], [38]) and then test whether the function is zero in a uniform sense ([8], [32]). In practice, however, it is often difficult to decide which series terms should be employed to approximate the unknown function: Using too few may induce bias, whereas using too many may not only distort the size of inference but also hurt its power. To address this issue, it seems natural to apply some machine learning based variable selection procedure such as the Lasso ([41]) or its variants. Although such methods may achieve the so-called "oracle property" in large samples, they cannot meet that theoretical ideal in finite samples. Ignoring the sampling variability in the selection step may thus lead to possibly severe size distortion in the subsequent test (see Section 3 for concrete Monte Carlo evidence).

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

We also characterize local alternatives against which the test is consistent. The power analysis clarifies a well-defined sense of adaptiveness of the proposed selective test: The test is able to detect smaller deviations from the null if the deviation has a simpler form. In the extreme case when the unknown function can be approximated by a bounded number of series terms (but with a priori unknown identities), the test achieves consistency nearly—up to a logarithmic factor—at the parametric rate. In the worst-case scenario in which the unknown function is "very complex" (e.g., all covariates have nontrivial predictive power), the power of the selective test deteriorates to the same level as the benchmark non-selective test. Outside the worst-case scenario, the proposed selective test is generally more powerful than the existing benchmark.

The finite-sample improvement afforded by the new test is demonstrated in a Monte Carlo study and an empirical application. Consistent with theory, the simulation results show that the selective test has excellent size control even in challenging inferential situations with small sample size and/or many candidate approximating functions; this is an important improvement over the existing non-selective test, as the latter can be severely over-sized under the same scenario. Moreover, the numerical results reveal that the selective test is much more efficient than the non-selective test judged by size-corrected power. The same benefits are also reflected in our empirical application on the conditional evaluation of inflation forecasts. In this real data setting, we document that the non-selective test can be very sensitive to the user's choice of the series approximation specification, which opens the door for potential "specification snooping," rendering the empirical findings hard to interpret. In contrast, the selective test delivers quite robust findings that are also sufficiently informative to help discriminate some popular prediction methods.

The remainder of this paper is organized as follows. We present the selective test and the related asymptotic theory in Section 2. Section 3 demonstrates the test's finite-sample performance in a Monte Carlo experiment. An empirical illustration on the conditional evaluation of inflation forecasts is provided in Section 4. Section 5 concludes. The appendix provides requisite implementation details. The Online Supplemental Appendix contains additional theoretical and simulation results, all technical proofs, and supplementary information for the empirical study.

The following notation will be adopted throughout the paper. All limits are for $n \to \infty$, with n denoting the sample size. We use $\|\cdot\|_\infty$, $\|\cdot\|$ and $\|\cdot\|_S$ to denote the matrix infinity norm, Frobenius norm and spectral norm, respectively. For two sequences of positive numbers a_n and b_n, we write $a_n \succ b_n$ if $a_n \geq c_n b_n$ for some strictly positive sequence $c_n \to \infty$.

2 A Selective Test for Conditional Moment Restrictions

2.1 The testing problem

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

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

with the associated residual term $\epsilon_t \equiv Y_t - g\left(X_t\right)$. The 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. This arises from many empirical economic settings. We consider a few examples to help fix ideas.

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

EXAMPLE 1 (CONDITIONAL EVALUATION OF PREDICTIVE ABILITY). Driven by the rapid development of modern data analytics, empirical analyses—especially those based on big data or machine learning techniques—often involve comparing the out-of-sample performance of predictive models. While the popular Diebold–Mariano test ([23]) pertains to the on-average or unconditional performance of predictive methods, [27] advocate a richer conditional evaluation framework. As a case in point, let Y_t be the differential of the ex post predictive losses of two competing methods at period t. Then the null hypothesis in (2.1) asserts that the methods have identical conditional performance across all conditioning states specified by X_t. As noted by [27], with Y_t properly defined, the test can also be used to test for conditional bias, rationality, and encompassing.

EXAMPLE 2 (UNCOVERED INTEREST RATE PARITY). Conditional moment restrictions in the form of (2.1) may also be implied by no arbitrage under rational expectation. For instance, a large literature in international finance has been devoted to testing the uncovered interest rate parity; see the seminal work of [26] and the reviews by [24, 25]. For such application, Y_t is the excess return on foreign bonds, defined as the exchange-rate-adjusted interest rate differential between foreign and domestic bonds. The X_t conditioning variable belongs to the investor's ex ante information set including, for example, the lagged interest rate differential.

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 arises from consumption-based asset pricing (see, e.g., [28]), in which the one-period-ahead pricing equation takes the form

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

where $U'\left(\cdot,\gamma\right)$ is the representative agent's marginal utility function with a preference parameter γ, β is the discounting factor, C_t is the consumption process, R_{t+1} is the return of an asset, and X_t is the state variable underlying the dynamic model. Equation (2.2) can be written in the form of (2.1) by setting $\theta^* = (\beta,\gamma)$ and $Y_t\left(\theta^*\right) = \frac{\beta U'(C_{t+1},\gamma)}{U'(C_t,\gamma)}R_{t+1}-1$. Similar equilibrium conditions can also be derived from Bellman equations; see [32] for an example in the context of a search-and-matching model for unemployment. In macroeconomic settings, the parameter θ^* is often, though not always, calibrated based on external data and auxiliary models.

The examples demonstrate the empirical relevance of the testing problem under study. Depending on the context, the Y_t variable may play distinct roles, possibly involving a finite-dimensional parameter θ^* that may be estimated or calibrated. In addition, it is generally important to accommodate time-series dependence in the data. For ease of exposition, we shall assume that Y_t is directly observed in this section. It is straightforward to further accommodate the presence of an unknown θ^*; the details are provided in Section SA of the supplemental appendix.

It is worth emphasizing that the hypothesis testing problem studied here is *functional* in nature, as it concerns the global property of the conditional expectation function $g(\cdot)$. In practice, empiricists often take "shortcuts" to bypass the functional inference problem, for example, by integrating out the conditioning variable X_t and simply testing the unconditional moment restriction $\mathbb{E}[Y_t] = 0$. To incorporate conditioning information, applied researchers tend 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. Evidently, if the null is violated in a way that is "orthogonal" to such parametric specification, the test would have little power in detecting it. Non-rejections may thus be challenged by a critical reader, because the "parametrized" test is designed to seek power only in specific directions that are generally hard to justify economically.

One solution that has long been advocated in econometrics is to adopt the series regression approach ([1], [38]) by including additional nonlinear approximating functions (e.g., polynomials, splines, trigonometric functions, wavelets) of X_t as explanatory variables, yielding

$$Y_t = b^\top P(X_t) + e_t, \tag{2.4}$$

where $P(\cdot)$ denotes the vector of said approximating functions. By letting the number of series terms grow to infinity, the specification becomes increasingly more flexible in larger samples and the series approximation approaches the true unknown function. The related functional testing problem has been studied in the recent literature on uniform nonparametric inference via growing dimensional Gaussian coupling techniques; see [8] and [32].

The existing econometric theory, however, does not provide clear guidance on how to specify the vector $P(\cdot)$ of approximating functions for nonparametric inference such as (2.1) in a general time series setting.[2] This is troublesome in practice because the specification may nontrivially influence the testing result especially when the sample size is not very large. The applied user may thus find it difficult to convince a skeptical audience that the chosen specification is not "cherry picked" to serve his empirical narrative.

To mitigate this concern, a common practice is to report robustness checks obtained from a range of alternative specifications. For instance, the user may conduct the test by fitting a pth order polynomial and demonstrate robustness by presenting results with different p's. This is easy to carry out and arguably adequate when the conditioning variable X_t is univariate. Unfortunately, this practice can easily lose meaningfulness when X_t is multivariate. To appreciate the practical difficulty, suppose that X_t is bivariate and the user employs bivariate polynomials to approximate the unknown function. Further imagine that the user aims to carry out a robustness check by varying p, say, from 4 to 8, which is adequately but not excessively wide. When $p = 4$, there are 15 series terms in (2.4). When $p = 8$, the number increases to 45. In a finite sample with moderate size, the latter relatively large number of regressors may severely distort inference because the underlying asymptotic argument does not "kick in" sufficiently well. Indeed, as we shall see from the simulation results in Section 3, implementing the test of [8] and [32] with "many" series terms will result in a substantial amount of false rejections. Robustness checks done in this way is then nothing more than an over-interpretation of type I

[2] Here, specifying $P(\cdot)$ refers not only to the dimension of $P(\cdot)$, but more importantly, to which conditioning variable(s) in X_t should enter $P(\cdot)$. In the literature, traditional data-driven methods for specifying $P(\cdot)$, such as cross-validation and Mallow's criterion (see, e.g., [33] and [3]), are proposed to determine the dimension of $P(\cdot)$ with a given order of components in $P(\cdot)$, to achieve optimal risk for the series estimator under the i.i.d. assumption. When applying the optimal tuning from these methods for inference, undersmoothing or bias correction are usually adopted. However, it is unclear if the randomness arising from the data-driven tuning should be formally accounted for when conducting inference, and if so, how to address this randomness.

errors. This makes the empiricist's task of defending any choice of the approximation function more challenging.

This motivates us to adopt a data-driven selection of the approximating functions for implementing the functional inference. The basic idea is simple. In a preliminary selection step, we first find approximating functions that are "sufficiently related" to the dependent variable Y_t by using a Lasso procedure. They are then included in (2.4) for the conduct of nonparametric inference. Under this framework, the user only needs to provide a "dictionary" of candidate approximating functions (such as a large collection of polynomials or splines), but will not dictate which of them enter the series regression (2.4). This—to a large extent—lessens the concern of "specification snooping."[3] Moreover, since the data-driven selection is able to discard many not-so-useful approximating functions, it also makes the statistical inference more efficient, resulting in a more powerful test. The proposed test based on data-driven selection will be referred to as the *selective test*.

Although the said construction is arguably straightforward to conceive, the requisite econometric analysis has a novel and nonstandard component. The main difficulty stems from the fact that the randomness in the selection step and that in the testing step are dependent. To carry out the selective test, we thus need to precisely characterize the said dependence so as to properly correct for the effect of the data-driven selection of series terms on the subsequent nonparametric test. As we shall detail in Section 2.3, the selection step exerts effect by restricting the distribution of certain influence function on a polytope. We correct for it accordingly by constructing critical values based on polytope-truncated Gaussian distributions, which are quite distinct from those relying on routine Gaussian approximations (cf. [8], [32]). Simulation results in Section 3 show clearly the benefit of the proposed correction.

We next turn to the details. Section 2.2 describes the selection procedure and the selective test statistic. Section 2.3 explains the effect of selection on the subsequent test and how to

[3] We clarify that the proposed method is not completely free of tuning, as it still requires the user to specify a dictionary and penalty parameters for the data-driven selection. However, the data-driven selection procedure enables a disciplined approach to choosing relevant regressors, avoiding ad hoc choices. Importantly, unlike methods that require regressors to be ordered a priori by their relevance (which is typically unknown), the Lasso-based selection does not depend on such an ordering. This differentiates our approach from earlier work focused on selecting the number of approximation terms based on an ordered list of series functions.

correct for it in the critical value construction. Section 2.4 presents theoretical properties of the proposed test.

2.2 The selective test statistic

Let $\mathcal{M} = \{1, \ldots, m\}$ and consider a collection $\{p_j(\cdot) : j \in \mathcal{M}\}$ 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 the proposed test statistics will be invariant to the ordering. We refer to $\mathcal{S}$ as a *selection* and denote its cardinality by $|\mathcal{S}|$. The series estimator for the conditional expectation function $g(\cdot)$ based on the selection $\mathcal{S}$ is given by

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

as long as the matrix $\widehat{Q}_{\mathcal{S}} \equiv n^{-1} \sum_{t=1}^{n} P_{\mathcal{S}}(X_t) P_{\mathcal{S}}(X_t)^{\top}$ is invertible. The standard error function associated with $\hat{g}_{\mathcal{S}}(\cdot)$ is

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

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 $\widehat{Q}_{\mathcal{S}}$ and estimate $\Sigma_{\mathcal{S}}$ using a (possibly growing dimensional) heteroskedasticity and autocorrelation consistent (HAC) estimator $\widehat{\Sigma}_{\mathcal{S}}$, following known results in the literature.[4] The standard error function $\sigma_{\mathcal{S}}(\cdot)$ can then be estimated via

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

[4] $\widehat{\Sigma}_{\mathcal{S}}$ may be taken as the classical Newey–West estimator or more generally the HAC estimators studied by [2]. The consistency and rate of convergence of these HAC estimators have been established in a general time-series setting with growing dimensions by [32]; 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 [19] and Lemma B2 in [32].

For a given selection $\mathcal{S}$, the associated sup-t statistic is defined as

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

The benchmark non-selective test studied by [8] and [32] relies on a special case of this test statistic with $\mathcal{S} = \mathcal{M}$.

Unlike [8] and [32], we aim to conduct the test on the basis of a data-driven selection. We use the Lasso method to implement the selection. The Lasso estimator is computationally attractive and has many desirable properties in both low- and high-dimensional settings, making it one of the most popular approaches for feature selection in the literature. Compared to classical selection methods based on information criteria, Lasso is more flexible as it treats all elements in the dictionary symmetrically, without requiring them to be ordered (which would simplify the selection as a choice of the number of series terms). Our approach may be extended to accommodate other types of variable selection methods as well, which we leave for future research.

Some users may consider a subset $\mathcal{M}_0 \subseteq \mathcal{M}$ of regressors to be important a priori and like to "manually" select them into the nonparametric fit. To accommodate this, 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 $\mathcal{M}_0^c \equiv \mathcal{M} \setminus \mathcal{M}_0$.[5] As such, Lasso assists the user's choice without dictating it completely.[6] 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.

The Lasso-assisted selection is implemented as follows. Given the user's prior choice $\mathcal{M}_0$,

[5] 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, she may put the constant and linear terms in $\mathcal{M}_0$, and let Lasso machine-learn whether and which additional terms are needed.

[6] It is worth pointing out that our goal is to conduct inference on the function $g(x)$, not on the parameters of the approximating functions indexed by $\mathcal{M}_0$. This makes our paper fundamentally different from another strand of research that focuses on the coefficients of a fixed set of regressors when there are many other regressors in linear regression (see, e.g., [15], [16] and [29] for recent developments in this line of research).

we perform a Lasso estimation with the resulting estimator given by

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

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 non-negative 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 $\omega_j = 1$ identically or $\omega_j = \|\mathbf{P}_j\|_n$ where $\|\mathbf{P}_j\|_n \equiv \sqrt{n^{-1} \sum_{t=1}^{n} p_j(X_t)^2}$ (see, e.g., [11] and [6]), but the more general setting in (2.8) also accommodates the adaptive Lasso ([44]). For the empirical implementation of the test proposed below, we use

$$\lambda_n = \sqrt{n \log(\log(n))} \Phi^{-1}(1 - 0.1/(2m)) \qquad \text{and} \qquad \omega_j = \hat{\sigma} \|\mathbf{P}_j\|_n, \tag{2.9}$$

where $\Phi^{-1}(\cdot)$ and $\hat{\sigma}$ denote the standard normal quantile function and a consistent estimator of the standard deviation of ϵ_t, respectively.[7] These values are computationally convenient and ensure that the proposed test satisfies a well-defined sense of adaptiveness, as elaborated in Section 2.4.[8] In applications, the ℓ_1-penalty tends to shrink many coefficients to zero. The data-driven selection is then given by

$$\mathcal{L} \equiv \mathcal{M}_0 \bigcup \left\{ j \in \mathcal{M}_0^c : \hat{\beta}_j^{Lasso} \neq 0 \right\}, \tag{2.10}$$

which consists of the user's ex ante choice $\mathcal{M}_0$ and Lasso's ex post selection from $\mathcal{M}_0^c$.

[7] In both the simulation study and the empirical application discussed in the paper, we construct $\hat{\sigma}^2$ using the sample variance of the estimated residuals from a Lasso estimation with penalty parameters $\lambda_n = \sqrt{n \log(\log(n))} \Phi^{-1}(1 - 0.1/(2m))$ and $\omega_j = \|\mathbf{P}_j\|_n$.

[8] We prefer the simple plug-in rule in (2.9) over data-driven ones for two reasons. First, the attractive properties of the Lasso estimator based on data-driven penalty parameters, such as those determined by cross-validation, are established for independent data (see, e.g., [20] and [22]). It is unclear if these properties are still valid in the time series setting. More importantly, data-driven penalty parameters introduce another source of randomness to the selective test, which may need to be accounted for to ensure valid size control.

The selective test statistic is defined accordingly as

$$\widehat{T}_{\mathcal{L}} \equiv \widehat{T}_{\mathcal{S}}\big|_{\mathcal{S}=\mathcal{L}} = \sup_{x\in\mathcal{X}} \left| \frac{n^{1/2}\hat{g}_{\mathcal{L}}(x)}{\widehat{\sigma}_{\mathcal{L}}(x)} \right|. \tag{2.11}$$

A large value of the test statistic signifies a violation of the null hypothesis (i.e., $g(\cdot) \neq 0$). It remains to properly determine the critical value, to which we now turn.

2.3 Critical value for the selective test

Before introducing our new critical value for the selective test statistic $\widehat{T}_{\mathcal{L}}$, it is instructive to briefly review how critical values are constructed in a simpler benchmark scenario with a non-random selection $\mathcal{S}$ as analyzed by [8] and [32]. The basic idea is to strongly approximate $\widehat{T}_{\mathcal{S}}$ by the supremum of a Gaussian process under the null hypothesis. More precisely, it can be shown under commonly used regularity conditions that there exists a sequence of $|\mathcal{S}|$-dimensional Gaussian random vectors $\widetilde{N}_{\mathcal{S}} \sim \mathcal{N}(0, \Sigma_{\mathcal{S}})$ such that

$$\widehat{T}_{\mathcal{S}} = \widetilde{T}_{\mathcal{S}} + o_p(1), \quad \text{where} \quad \widetilde{T}_{\mathcal{S}} \equiv \sup_{x\in\mathcal{X}} \left| \frac{P_{\mathcal{S}}(x)^{\top} Q_{\mathcal{S}}^{-1} \widetilde{N}_{\mathcal{S}}}{\sigma_{\mathcal{S}}(x)} \right|. \tag{2.12}$$

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 via simulation as the $1-\alpha$ quantile of

$$\widetilde{T}^{*}_{\mathcal{S}} \equiv \sup_{x\in\mathcal{X}} \left| \frac{P_{\mathcal{S}}(x)^{\top} \widehat{Q}_{\mathcal{S}}^{-1} \widetilde{N}^{*}_{\mathcal{S}}}{\widehat{\sigma}_{\mathcal{S}}(x)} \right|, \tag{2.13}$$

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

From here, a seemingly natural way to construct $\widehat{T}_{\mathcal{L}}$'s critical value is to directly apply the same procedure by plugging in $\mathcal{S} = \mathcal{L}$, which amounts to ignoring the randomness in the data-driven selection $\mathcal{L}$. However, this approach turns out to suffer nontrivial size distortion as shown in the simulation study in Section 3. This motivates us to account for the effect of selection and adjust the critical value accordingly. This is in fact the key ingredient that makes our proposed selective test work reliably in finite samples.

Since the formal analysis is technical, we summarize the main intuition in the next two paragraphs before diving into the econometrics. We begin with a couple of clarifications. First, while the reason for performing the data-driven selection of "relevant" approximating functions is to fit the unknown function $g(\cdot)$ under the alternative hypothesis, the need for correcting the selection effect stems from our desire to control size under the null hypothesis. Hence, the discussion in the remainder of this subsection concentrates on the null hypothesis, while taking the selection algorithm as given. Second, although the Lasso regression is informative about which approximating function in the dictionary is useful for fitting the unknown $g(\cdot)$ function, it is silent on whether one should reject the hypothesis (2.1) or not, simply because it is not designed as a test. In particular, the fact that some approximating functions are selected by Lasso does not mean—in a formal sense of hypothesis testing—that the null hypothesis needs to be rejected.

Why the data-driven selection may lead to size distortion? Note that under the null hypothesis, all approximating functions in the dictionary are useless for fitting $g(\cdot)$, which is identically zero. Ideally, Lasso should select nothing. But it is imperfect in reality. In finite samples, Lasso will sometimes select approximating functions which *appear* (counterfactually) useful for fitting Y_t purely due to random disturbances. When taking the selection "as given," the empiricist is implicitly conditioning on an "unusual" random event (associated with the selection outcome). Since the selection decision and the subsequent test are based on "correlated" information, the conditioning will alter the distribution of the test statistic, rendering the usual asymptotic Gaussian approximation inaccurate in finite samples; see Section 3 for concrete numerical evidence. To address this issue, we shall explicitly characterize the selection event in the form of a collection of inequality constraints. This allows us to determine the joint sampling behavior of the data-driven selection and the selective test statistic, so that we can compute critical values properly by taking into account the said selection-induced conditioning effect.[9]

We now proceed to the details. The technical discussion in the reminder of this subsection,

[9] Alternatively, one may apply sample splitting or cross-fitting to reduce the size distortion introduced by model selection. However, formal justification of these methods typically requires the data to be independent. Since our primary applications are in the time series setting, these methods are not discussed in the paper, and their theoretical justification is left for future investigation.

together with the asymptotic theory in Section 2.4, may be skipped by readers who are mainly interested in applications. 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 $\widetilde{\mathbf{P}}_{\mathcal{S}}$ and $\widetilde{\mathbf{G}}$ represent the residual matrices derived from projecting $\mathbf{P}_{\mathcal{S}}$ onto $\mathbf{P}_{\mathcal{M}_0}$ and $\mathbf{G}$ onto $\mathbf{P}_{\mathcal{M}_0}$, respectively. In other words, $\widetilde{\mathbf{P}}_{\mathcal{S}} \equiv \mathbf{D}_n\mathbf{P}_{\mathcal{S}}$ and $\widetilde{\mathbf{G}} \equiv \mathbf{D}_n\mathbf{G}$, 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 $\hat{\mathrm{s}}$ 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.[10] Specifically, for any nonrandom selection $\mathcal{S}$ satisfying $\mathcal{M}_0 \subseteq \mathcal{S} \subseteq \mathcal{M}$ and a sign vector $\mathrm{s} \in \{\pm 1\}^{|\mathcal{S}\setminus\mathcal{M}_0|}$, we have

$$\{\mathcal{L} = \mathcal{S}, \hat{\mathrm{s}} = \mathrm{s}\} = \left\{n^{-1/2}\mathbf{P}^\top\boldsymbol{\epsilon} \in \Pi(\mathcal{S}, \mathrm{s}, \lambda_n)\right\}. \tag{2.14}$$

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

$$\Pi(\mathcal{S}, \mathrm{s}, \lambda_n) \equiv \left\{ z \in \mathbb{R}^m : \begin{array}{c} \mathrm{diag}(\mathrm{s})\,(A_{\mathcal{S}}z + c_{\mathcal{S}}) > n^{-1/2}\lambda_n b_{\mathcal{S}}(\mathrm{s}) \text{ and} \\ \\ n^{-1/2}\lambda_n b'_{l,\mathcal{S}}(\mathrm{s}) < A'_{\mathcal{S}}z + c'_{\mathcal{S}} < n^{-1/2}\lambda_n b'_{u,\mathcal{S}}(\mathrm{s}) \end{array} \right\}, \tag{2.15}$$

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

$$\left\{ \begin{array}{lcl} c_{\mathcal{S}} & \equiv & n^{1/2}\left(\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^\top\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}\right)^{-1}\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^\top\widetilde{\mathbf{G}}, \\ \\ c'_{\mathcal{S}} & \equiv & n^{-1/2}\widetilde{\mathbf{P}}_{\mathcal{M}\setminus\mathcal{S}}^\top\left(\mathbf{I}_n - \widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}\left(\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^\top\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}\right)^{-1}\widetilde{\mathbf{P}}_{\mathcal{S}\setminus\mathcal{M}_0}^\top\right)\widetilde{\mathbf{G}}, \end{array} \right. \tag{2.16}$$

[10] See Lemma SA.1 in the Supplemental Appendix for details, which extends a similar result in [31] by allowing for the prior choice set $\mathcal{M}_0$ and penalty weights ω_j.

and $b_{\mathcal{S}}(\mathrm{s})$, $b'_{l,\mathcal{S}}(\mathrm{s})$, $b'_{u,\mathcal{S}}(\mathrm{s})$, $A_{\mathcal{S}}$, and $A'_{\mathcal{S}}$ are directly observable quantities. These observed statistics do not pose 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 $\widetilde{\mathbf{G}}$ involves the unknown $g(\cdot)$ function. The structure of the polytope $\Pi(\mathcal{S}, \mathrm{s}, \lambda_n)$ is not directly observed, either.

The above (non-asymptotic) characterization precisely depicts the relation between the selection and the subsequent series estimation in finite samples, through their common dependence on the score vector $n^{-1/2}\mathbf{P}^{\top}\boldsymbol{\epsilon}$. To see this more clearly, recall that for any given selection $\mathcal{S}$, the asymptotic normality (formulated in terms of strong Gaussian coupling in the growing dimensional case) 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 s, the score $n^{-1/2}\mathbf{P}^{\top}\boldsymbol{\epsilon}$ is restricted within the polytope $\Pi(\mathcal{S}, \mathrm{s}, \lambda_n)$. This restriction would modify the score's asymptotic normality into a form of truncated normality. A failure to account for this will generally lead to size distortion.[11]

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 $\Pi(\mathcal{S}, \mathrm{s}, \lambda_n)$. 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

$$\widehat{b}_{\mathcal{S}} \equiv \left(\mathbf{P}_{\mathcal{S}}^{\top}\mathbf{P}_{\mathcal{S}}\right)^{-1}\mathbf{P}_{\mathcal{S}}^{\top}\mathbf{Y}. \tag{2.17}$$

We further modify this preliminary estimator via hard-thresholding to obtain $\tilde{\beta}_{\mathcal{S}}$, with its jth component given by

$$\tilde{\beta}_{\mathcal{S},j} \equiv \widehat{b}_{\mathcal{S},j} \cdot 1\left\{|\widehat{b}_{\mathcal{S},j}| \geq \log(n)n^{-1/2}\widehat{\sigma}_{\mathcal{S},j}\right\}, \tag{2.18}$$

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 $\widehat{b}_{\mathcal{S},j}$

[11] It is worth noting that the said distortion is distinct from the usual small-sample phenomenon that central limit theorems may not "kick in" sufficiently well for a moderately sized sample; indeed, the same issue still arises even if the score is exactly normally distributed (say, in a Gaussian model with fixed design).

obtained as the square-root of the jth diagonal element of $\widehat{Q}_{\mathcal{S}}^{-1}\widehat{\Sigma}_{\mathcal{S}}\widehat{Q}_{\mathcal{S}}^{-1}$.[12] The n-dimensional vector $\mathbf{G}$ is then approximated by $\mathbf{P}_{\mathcal{S}}\tilde{\beta}_{\mathcal{S}}$. Plugging this approximation into (2.16), we further obtain approximations for $c_{\mathcal{S}}$ and $c'_{\mathcal{S}}$ in the form of

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

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 $\Pi(\mathcal{S}, \mathrm{s}, \lambda_n)$ can then be obtained by replacing $(c_{\mathcal{S}}, c'_{\mathcal{S}})$ with $(\widehat{c}_{\mathcal{S}}, \widehat{c}'_{\mathcal{S}})$ in (2.15), that is,

$$\widehat{\Pi}(\mathcal{S}, \mathbf{s}, \lambda_n) \equiv \left\{ z \in \mathbb{R}^m : \begin{array}{c} \operatorname{diag}(\mathbf{s})\,(A_{\mathcal{S}}z + n^{1/2}\tilde{\beta}_{\mathcal{S}\setminus\mathcal{M}_0}) > n^{-1/2}\lambda_n b_{\mathcal{S}}(\mathbf{s}) \text{ and} \\ \\ n^{-1/2}\lambda_n b'_{l,\mathcal{S}}(\mathbf{s}) < A'_{\mathcal{S}}z < n^{-1/2}\lambda_n b'_{u,\mathcal{S}}(\mathbf{s}) \end{array} \right\}. \tag{2.19}$$

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 $\widehat{\Sigma}_{\mathcal{M}}^{1/2}\widetilde{N}^*$ extracted in accordance with $\mathcal{S}$ as a subset of $\mathcal{M}$, and use it to compute $\widetilde{T}^*_{\mathcal{S}}$ as described in (2.13). We then set

$$cv^1_{\mathcal{S},\alpha} \equiv \inf\left\{ u \in \mathbb{R} : \frac{\mathbb{P}^*\left(\widetilde{T}^*_{\mathcal{S}} \geq u, \widehat{\Sigma}_{\mathcal{M}}^{1/2}\widetilde{N}^* \in \widehat{\Pi}(\mathcal{S}, \mathbf{s}, \lambda_n)\right)}{\mathbb{P}^*\left(\widehat{\Sigma}_{\mathcal{M}}^{1/2}\widetilde{N}^* \in \widehat{\Pi}(\mathcal{S}, \mathbf{s}, \lambda_n)\right)} = \alpha \right\}, \tag{2.20}$$

where $\mathbb{P}^*(\cdot)$ denotes the conditional distribution of $\widetilde{N}^*$ given data.[13,14] Our proposed critical

[12] The intuition for applying the hard-thresholding 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.

[13] 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 (see, e.g., [12]).

[14] If the focus is on testing $g(x_0) = 0$ for some given $x_0 \in \mathcal{X}$, then the corresponding critical value may be calculated using techniques from recent strand of literature on "selective inference" (see, e.g.,

value takes a hybrid form of $cv^0_{\mathcal{L},\alpha} \equiv cv^0_{\mathcal{S},\alpha}\big|_{\mathcal{S}=\mathcal{L}}$ and $cv^1_{\mathcal{L},\alpha} \equiv cv^1_{\mathcal{S},\alpha}\big|_{\mathcal{S}=\mathcal{L}}$, that is,

$$cv_{\mathcal{L},\alpha} \equiv cv^0_{\mathcal{L},\alpha} + (cv^1_{\mathcal{L},\alpha} - cv^0_{\mathcal{L},\alpha}) \cdot 1\{\widehat{T}_{\mathcal{L}} \leq \log(n)\}. \tag{2.21}$$

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 $\{n^{-1/2}\mathbf{P}^\top\boldsymbol{\epsilon} \in \Pi(\mathcal{S}, \mathrm{s}, \lambda_n)\}$ is captured by that of $\widetilde{T}^*_{\mathcal{S}}$ and $\{\widehat{\Sigma}^{1/2}_{\mathcal{M}}\widetilde{N}^* \in \widehat{\Pi}(\mathcal{S}, \mathrm{s}, \lambda_n)\}$. The critical value $cv^1_{\mathcal{S},\alpha}$ described in (2.20) is simply defined as a tail quantile of the conditional distribution of $\widetilde{T}^*_{\mathcal{S}}$ in restriction to the "coupling" selection event $\{\widehat{\Sigma}^{1/2}_{\mathcal{M}}\widetilde{N}^* \in \widehat{\Pi}(\mathcal{S}, \mathrm{s}, \lambda_n)\}$, which captures how the polytope restriction on the score vector distorts the distribution of the sup-t statistic. Although $cv^1_{\mathcal{L},\alpha}$ is preferred for size control, under the alternative hypothesis, we may use $cv^0_{\mathcal{L},\alpha}$ for inference since its magnitude is small (at most $\log(m)^{1/2}$ under some regularity conditions). The indicator $1\{\widehat{T}_{\mathcal{L}} \leq \log(n)\}$ in (2.20) defines a switching rule between $cv^0_{\mathcal{L},\alpha}$ and $cv^1_{\mathcal{L},\alpha}$. With probability approaching 1, this indicator equals 1 under the null, and 0 under the alternatives specified in (2.24) and (2.28) below. This ensures both good size and power properties for the inference based on $cv_{\mathcal{L},\alpha}$.

We close this subsection with a few remarks. 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 ([42]) and the elastic net ([45], [46]), one may modify the underlying Karush–Kuhn–Tucker conditions accordingly and characterize the selection event in a similar fashion as (2.14). Critical values may then be constructed from the corresponding conditional coupling distributions. Secondly, we stress that our analysis focuses on testing whether $g(\cdot) = 0$, and hence, our inference concentrates on the null hypothesis. A separate open question is how to

[31]). However, our focus is on uniform inference of $g(x)$ as specified in (2.1), which makes these techniques not applicable here. That being said, selective inference may be fruitfully used in many other econometric problems, as demonstrated in the recent interesting work by [34], [4, 5]. Also see [35] for a novel power-improved approach to selective inference problems.

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

2.4 Asymptotic properties of the selective test

In this subsection, we 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. We focus on the baseline setting in which Y_t is directly observed. Section SA in the supplemental appendix details an extension with Y_t depending on some unknown parameter θ^*. We start with introducing a few regularity conditions.

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

Assumption 1. *(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 1 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., [18]). The 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 $\mathcal{M}_0^c$ with nonzero loadings. Note that $\mathcal{R}$ is empty under the null hypothesis, but it plays an important role under the alternative.

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

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

which is arguably the best one may wish to obtain from any selection algorithm. 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.

Assumption 2. *(i) The eigenvalues of $Q_{\mathcal{M}}$ and $\Sigma_{\mathcal{M}}$ are bounded from above and away from zero; (ii) $\sup_{\gamma\in\mathcal{B}_m^\star} |\gamma^\top(\widehat{Q}_{\mathcal{M}} - Q_{\mathcal{M}})\gamma| = o_p(1)$, where $\mathcal{B}_m^\star \equiv \{\gamma \in \mathbb{R}^m\colon ||\gamma||_0 \leq 2|\mathcal{M}^\star|\log(n), ||\gamma|| = 1\}$; (iii) $(||A'_{\mathcal{M}_0}\mathbf{P}^\top\boldsymbol{\epsilon}||_\infty + (n\log(m))^{1/2})/(\lambda_n \min_{j\in\mathcal{M}_0^c\setminus\mathcal{R}}\omega_j) = o_p(1)$; (iv) there exists a sequence of $|\mathcal{M}_0|$-dimensional Gaussian random vectors $\widetilde{N}_{\mathcal{M}_0}$ with zero mean and variance $\Sigma_{\mathcal{M}_0}$ such that*

$$n^{-1/2}\sum_{t=1}^{n} P_{\mathcal{M}_0}(X_t)\epsilon_t = \widetilde{N}_{\mathcal{M}_0} + o_p(\log(n)^{-1});$$

(v) the estimator $\widehat{\Sigma}_{\mathcal{M}}$ of $\Sigma_{\mathcal{M}}$ satisfies $\sup_{\gamma\in\mathcal{B}_m^\star}\left|\gamma^\top(\widehat{\Sigma}_{\mathcal{M}} - \Sigma_{\mathcal{M}})\gamma\right| = o_p(1)$; (vi) $||\widehat{Q}_{\mathcal{M}_0} - Q_{\mathcal{M}_0}||_S + ||\widehat{\Sigma}_{\mathcal{M}_0} - \Sigma_{\mathcal{M}_0}||_S = o_p(|\mathcal{M}_0|^{-1/2}\log(n)^{-3/2})$; (vii) $\log(\zeta_m^L) = O(\log(m))$ and $\log(m) = O(\log(n))$, where $\zeta_m^L \equiv \sup_{x,x'\in\mathcal{X}} \|P(x) - P(x')\| \,/\, \|x - x'\|$.

Assumption 2 imposes high-level conditions that are similar to those used in prior work on uniform series-based inference and Lasso estimation of high-dimensional models.[15] Condition (i) is fairly standard for series estimation; see, for example, [1], [38], and [18]. Conditions (ii, v) entail restricted matrix law of large numbers on $\widehat{Q}_{\mathcal{M}}$ and $\widehat{\Sigma}_{\mathcal{M}}$ respectively. These conditions impose restriction on the dimension of $\mathcal{M}^{\star}$ and are commonly used in the literature for estimation and inference of high-dimensional models (see, e.g., [11], [6] and [9]). Conditions (ii, v), combined with Condition (i), implies that

$$K^{-1} \leq \gamma^{\top} \widehat{Q}_{\mathcal{M}} \gamma \leq K \text{ and } K^{-1} \leq \gamma^{\top} \widehat{\Sigma}_{\mathcal{M}} \gamma \leq K, \tag{2.22}$$

uniformly over $\gamma \in \mathcal{B}_m^{\star}$, with probability approaching 1 (wpa1). Condition (iii) places a restriction on the penalty parameters λ_n and $\{\omega_j\}_{j \in \mathcal{M}_0^c \setminus \mathcal{R}}$ to ensure desirable statistical properties for the Lasso estimator. In many scenarios, we can demonstrate that $\left\| A'_{\mathcal{M}_0} \mathbf{P}^{\top} \boldsymbol{\epsilon} \right\|_{\infty} = O_p(n \log(m)^{1/2})$.[16] Hence, this condition holds under these circumstances if

$$n^{1/2} \log(m)^{1/2} (\lambda_n \min_{j \in \mathcal{M}_0^c \setminus \mathcal{R}} \omega_j)^{-1} = o_p(1).$$

The latter condition has been extensively discussed in literature such as [11], [6] and [21]. Condition (iv) requires that the scaled score $n^{-1/2} \sum_{t=1}^{n} P_{\mathcal{M}_0}(X_t) \epsilon_t$ admits a Gaussian coupling, which may be verified by applying Yurinskii's coupling for i.i.d. data (see, e.g., [7] and [14]) or for martingale (see, e.g., [17]), or the theory of [32] in the more general time-series setting for heterogeneous mixingales. Condition (vi) pertains to the convergence rates of $\widehat{Q}_{\mathcal{M}_0}$ and $\widehat{\Sigma}_{\mathcal{M}_0}$, which can be verified under primitive conditions as shown in [19] and [32]. Remarkably, Conditions (iv, vi) are imposed on the user-specified set $\mathcal{M}_0$, which might have a fixed or small dimension. Condition (vii) is satisfied by commonly employed series bases.

Assumption 3. *For some fixed constant K_ω, $K_\omega^{-1} \leq \min_{j \in \mathcal{M}_0^c} \omega_j \leq \max_{j \in \mathcal{M}_0^c} \omega_j \leq K_\omega$ wpa1.*

Assumption 3 sets upper and lower bounds on individual penalties $\{\omega_j\}_{j \in \mathcal{M}_0^c}$. This condition can be verified for various choices, including the recommended one in (2.9) (see Proposition 1)

[15] A detailed discussion on verifying the conditions in Assumption 2 is included in Section SC of the Supplemental Appendix.

[16] See Lemmas SC.15 and SC.16 in Section SC of the Supplemental Appendix.

in the Appendix.

We are now ready to state the asymptotic size and power properties of the selective test, which is the main theoretical result of this paper. Below, let $\kappa\left(\mathcal{S}\right) \equiv \sup_{x\in\mathcal{X}} ||P_{\mathcal{S}}(x)||$ for any $\mathcal{S} \subset \mathcal{M}$, $m_0 \equiv |\mathcal{M}_0| + 1$ and $\mu_{\mathcal{R},m} \equiv |\mathcal{R}|^{1/2}\xi_m^{1/2}$, where ξ_m is defined in (2.24) below.

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) Suppose that there exists a non-random sequence ξ_m such that

$$\max_{\mathcal{S}\subset\mathcal{M}_0^c} \kappa\left(\mathcal{S}\right)^2 \left|\mathcal{S}\right|^{-1} \leq \xi_m, \tag{2.23}$$

then the selective test is consistent against any local alternative that satisfies

$$\sup_{x\in\mathcal{X}} |g_n(x)| \succ \frac{(\kappa(\mathcal{M}^\star) + \mu_{\mathcal{R},m})((1+\mu_{\mathcal{R},m})\sqrt{\lambda_n^2|\mathcal{R}|n^{-1} + \log(m_0)} + \log(n))}{n^{1/2}}, \tag{2.24}$$

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 (2.24), with the "boundary" of the local neighborhood under the uniform metric characterized by the rate in the right hand side of the inequality in (2.24). Condition (2.23) imposes a uniform upper bound ξ_m on the vector of approximating functions $P_{\mathcal{S}}(x)$. For approximating functions such as splines and wavelets, which are uniformly bounded, (2.23) holds with ξ_m being a fixed constant. For polynomials, (2.23) holds with $\xi_m = Km^{1/2}$ for some fixed K.

From the local alternatives in (2.24), it is evident that a large penalty parameter λ_n will increase the boundary rate, thereby reducing the power of the selective test. Conversely, λ_n should not be too small, as this may cause Assumption 2(iii) to be violated. This trade-off informs our recommended λ_n in (2.9), under which the boundary rate in (2.24) becomes

$$n^{-1/2}(\kappa(\mathcal{M}^{\star}) + \mu_{\mathcal{R},m})(1 + \mu_{\mathcal{R},m}) \log(n). \tag{2.25}$$

We will next discuss this rate in different scenarios, since it pertains to the recommended λ_n.

First, we compare the selective test against the non-selective test based on the user-specified set $\mathcal{M}_0$, where the latter rejects the null hypothesis (2.1) if $\widehat{T}_{\mathcal{M}_0} > cv^0_{\mathcal{M}_0,\alpha}$. The most favorable case for the non-selective test (but the least favorable to the selective test) is when $\mathcal{M}_0$ is oracle, meaning $\mathcal{R} = \varnothing$ and $\mathcal{M}^{\star} = \mathcal{M}_0$ and hence $\mathcal{R} = \varnothing$. In this case, we can show that the test statistic based on $\widehat{T}_{\mathcal{M}_0}$ is consistent against any local alternative that satisfies

$$\sup_{x \in \mathcal{X}} |g_n(x)| \succ n^{-1/2} \kappa\left(\mathcal{M}_0\right) \log(m_0)^{1/2}. \tag{2.26}$$

On the other hand, since $\mathcal{R} = \varnothing$ and $\mathcal{M}^{\star} = \mathcal{M}_0$ in this case, the rate in (2.25) simplifies to $n^{-1/2}\kappa\left(\mathcal{M}_0\right)\log(n)$. Therefore, the cost that the selective test incurs compared with the non-selective test for not knowing $\mathcal{M}^{\star} = \mathcal{M}_0$ is $\log(n)/\log(m_0)^{1/2}$, which is not substantial.

More importantly, Theorem 1(b) shows that the selective test is adaptive with respect to the complexity of $g_n(\cdot)$ as gauged by $\mathcal{R}$. 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., $\mathcal{R}$ is smaller), despite the fact that this information is unknown a priori. This is an important improvement relative to the existing method, which employs the user-specified set $\mathcal{M}_0$ or the entire dictionary $\mathcal{M}$ of basis functions. The power of the non-selective test based on $\mathcal{M}$ is always dictated by the fast-diverging sequence m, and hence low, regardless of the actual complexity underlying the data generating process. Meanwhile, when $\mathcal{R}$ is finite the selective test can be consistent at nearly (up to a logarithmic factor) parametric rate.

It is also remarkable that the consistency of the selective test stated in Theorem 1(b) is achieved without specifying the magnitudes of the coefficients b^*_n in the series approximation in Assumption 1(i). Therefore, b^*_n could be either hard-sparse, meaning $\min_{j\in\mathcal{R}} |b^*_{n,j}|$ is bounded away from zero, or approximate-sparse, meaning $\max_{j\in\mathcal{R}} |b^*_{n,j}|$ approaches zero. On the other hand, it is well-known that Lasso is generally inconsistent in model selection under Assumption 3, even when b_n^* is hard-sparse (see related discussion in [43] and [44]), which leads to the presence of $\mu_{\mathcal{R},m}$ in the boundary rate of the local alternatives.

When b_n^* is hard-sparse or approximate-sparse but $\min_{j\in\mathcal{R}}|b_{n,j}^*|$ decreases at a rate slower than $n^{-1/2}$, the set $\mathcal{R}$ can be consistently estimated through Lasso with adaptive individual penalties $\{\omega_j\}_{j\in\mathcal{M}_0^c}$, whose properties are specified in Assumption 4 below. In this case, condition (2.23) is not required, and a lower bound sharper than (2.24) on the local alternatives against which the selective test is consistent can be obtained.

Assumption 4. *The penalty parameters $\{\omega_j\}_{j\in\mathcal{M}_0^c}$ satisfy:*

$$\frac{|\mathcal{M}^\star|^{1/2}n^{1/2}+|\mathcal{R}|^{1/2}\lambda_n\max_{j\in\mathcal{R}}\omega_j}{\min\left\{n\min_{j\in\mathcal{R}}|b_{n,j}^*|,\lambda_n\min_{j\in\mathcal{M}_0^c\setminus\mathcal{R}}\omega_j\right\}}=o_p(1). \tag{2.27}$$

To understand how Assumption 4 ensures consistent estimation of $\mathcal{R}$, we first observe that $|\mathcal{M}^\star|^{1/2}n^{-1/2}+\lambda_n\max_{j\in\mathcal{R}}\omega_j|\mathcal{R}|^{1/2}n^{-1}$ represents the convergence rate of the Lasso estimator $\hat{\beta}^{Lasso}$. The signs of $b_{n,\mathcal{R}}^*$ are consistently estimated if $\min_{j\in\mathcal{R}}|b_{n,j}^*|$ dominates the estimation error. This means Assumption 4 holds when the denominator in the ratio of (2.27) becomes $n\min_{j\in\mathcal{R}}|b_{n,j}^*|$. On the other hand, the zero subvector $b_{n,\mathcal{M}_0^c\setminus\mathcal{R}}^*$ of b_n^* is estimated as zero in Lasso estimation if the penalty $n^{-1}\lambda_n\min_{j\in\mathcal{M}_0^c\setminus\mathcal{R}}\omega_j$ dominates the estimation error of $\hat{\beta}^{Lasso}$. This means Assumption 4 holds when the denominator in the ratio of (2.27) becomes $\lambda_n\min_{j\in\mathcal{M}_0^c\setminus\mathcal{R}}\omega_j$. Therefore, these conditions ensure that $\text{sign}(\hat{\beta}_{\mathcal{M}_0^c}^{Lasso})=\text{sign}(b_{n,\mathcal{M}_0^c}^*)$ with probability approaching 1, and consistent estimation of $\mathcal{R}$ can be achieved.

It should be noted that the recommended penalty parameters in (2.9) do not satisfy Assumption 4. In the Appendix of the paper, we provide an algorithm for obtaining penalty parameters that satisfy this assumption.

Theorem 2. *Suppose that Assumptions 1, 2, and 4 hold. The selective test has asymptotic level α for any $\alpha\in(0,1/2)$ under the null hypothesis (2.1). Moreover it is consistent against any local alternative that satisfies*

$$\sup_{x\in\mathcal{X}}|g_n(x)|\succ\kappa\left(\mathcal{M}^\star\right)(|\mathcal{M}^\star|^{1/2}+\log(n))n^{-1/2}, \tag{2.28}$$

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

Theorem 2 again demonstrates that the selective test adapts to the latent complex structure of the unknown function $g_n(x)$ measured by $\mathcal{M}^{\star}$. The rate in (2.28) depends solely on the unknown "oracle" selection $\mathcal{M}^{\star}$, which is achieved through consistent estimation of $\mathcal{R}$ in Lasso estimation under (2.27). The term $|\mathcal{M}^{\star}|^{1/2}$ in (2.28) can be replaced by $\log(1+|\mathcal{M}^{\star}|)^{1/2}$, if Assumption 2(iv) holds with $\mathcal{M}_0$ replaced by $\mathcal{M}^{\star}$. In this scenario, Theorem 2 can be further refined to show that the selective test is consistent against local alternatives that decay to zero at a rate slower than $\kappa\left(\mathcal{M}^{\star}\right)\log(n)n^{-1/2}$.

The selective test can also be applied to test other features, such as the derivatives of $g_n(x)$, with some modifications to the test statistic and the critical value. For example, suppose x is univariate and the research interest is in testing $\partial g_n(x)/\partial x = 0$ for all $x \in \mathcal{X}$. In this case, the test statistic and the critical value are defined similarly to $\widehat{T}_{\mathcal{L}}$ and $cv_{\mathcal{L},\alpha}$, with $P_{\mathcal{L}}(x)$ in the latter replaced by $\partial P_{\mathcal{L}}(x)/\partial x$. Assumptions 1 and 2(vii) shall hold for $\partial g_n(x)/\partial x$ and $\tilde{\zeta}_m^L \equiv \sup_{x,x'\in\mathcal{X}} ||\partial P(x)/\partial x - \partial P(x')/\partial x||/||x - x'||$ respectively, to ensure that the main results presented in Theorems 1 and 2 still apply.

Consistent with the theory, simulation results presented below show that the proposed selective test not only controls size much better than the existing non-selective test, but is also notably more powerful. We next turn to the details.

3 Monte Carlo Simulations

3.1 The simulation setting

We examine the finite-sample size and power properties of the proposed test using the following data generating process (DGP). 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 autocorrelation. 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 $\delta \neq 0$, the DGP is under the alternative hypothesis and δ quantifies how far the alternative deviates from the null. Below, we set $\delta = 0$ for the size analysis and use $\delta \in \{0.1, 0.2, \ldots, 1\}$ to trace out a test's power curve. The sample size is set as $n = 150$, 250, or 500, which is empirically relevant for typical time series applications. The number of Monte Carlo replications is 10,000.

To implement the selective test, we choose the Lasso penalty parameters according to (2.9), and then implement the test as described in Section 2.2. The prior choice set $\mathcal{M}_0$ contains only the constant term, which is our default recommendation. For comparison, we also consider two other tests. The first is the non-selective test of [8] and [32], 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 (2.13). 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 use cubic splines to generate bivariate approximating functions. In particular, we transform $X_{j,t}$ onto $[0, 1]$ using its empirical cumulative distribution function (here calibrated to a normal distribution) and then rescale it linearly onto the $[-1, 1]$ interval. The univariate cubic spline functions are given by

$$\mathscr{L}_j\left(x\right) = \begin{cases} x^{j-1}, & \text{if } j = 1, \ldots, 4 \\ \max(x - t_{j-4}, 0)^3, & \text{if } j > 4 \end{cases},$$

where the knots $t_1, ..., t_{j-4}$ are equally spaced between $[-1, 1]$. For any integer $m' \geq 2$, we generate $m = m'(m'-1)/2$ bivariate series functions by collecting $\mathscr{L}_{j_1}(x_1)\mathscr{L}_{j_2}(x_2)$ for $j_2, j_2 \geq 1$ and $j_2 + j_2 \leq m'$.

Finally, in order to examine how the finite-sample performance of the tests depends on the pre-specified dictionary $\mathcal{M}$, we consider $m' = 4$, 6, 8 and 10; the corresponding dictionary $\mathcal{M}$ contains $m = 6$, 15, 28 and 45 terms, respectively.

3.2 Simulation 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.

[Table 1 Here]

Panel A of Table 1 shows that the proposed 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 in all specifications of the sample size and the number of approximating functions in the dictionary. The results for the non-selective test, reported on Panel B, show a sharp contrast. When $m = 15$, the non-selective test shows some nontrivial over-rejection (with 19.4% rejection rate) when $n = 150$, though this is a small-sample phenomenon, because the size distortion diminishes quickly as we increase the sample size to $n = 500$, consistent with the asymptotic theory of [8] and [32]. However, the over-rejection becomes substantially more severe for larger

$\mathcal{M}$. Indeed, when $m = 45$, the non-selective test mistakenly rejects the null hypothesis more then 60% of the time when the sample size $n = 150$, and the rejection rate is still above 40% 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 does not provide an adequate finite-sample approximation when the dimension grows "too fast" relative to the sample size. This does not appear to be an issue for the selective test, because the data-driven 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 is 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. That being said, it is important to observe that the uncorrected test often over-rejects by a nontrivial amount when the sample size is relatively small, and so, is 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 demonstrates the necessity of accounting for the truncation effect induced by the data-driven selection as we have discussed in Section 2.2.

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 severe size distortion in general. The selective test is clearly the most reliable method among the three.

Next, we compare the finite-sample power of these tests. For brevity, we focus on the setting with $n = 500$. Since the non-selective and uncorrected tests generally suffer nontrivial size distortions, directly comparing their power with that of the selective test is problematic, as the most size-distorted test may (misleadingly) appear to be the most powerful. We thus focus on the size-adjusted power instead. Figure 1 plots the size-adjusted power curves for the selective, non-selective, and uncorrected tests for $m = 15$ and 45.[17]

[Figure 1 Here]

Looking at Figure 1, we first note that the size-adjusted power curves of all three tests hit the 5% nominal level at $\delta = 0$ by construction and, as expected, their rejection rates are increasing in δ. From the top row of the figure, we see that the proposed selective test and the uncorrected test have similar power properties when $m = 15$, and they are more powerful than the benchmark non-selective test. The latter finding is consistent with the intuition that the Lasso-assisted selection helps the tests seek power in a targeted fashion.

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

In summary, the simulation study shows that the proposed selective test has excellent size control across a broad range of scenarios, and is notably more powerful than the non-selective test. These findings clearly demonstrate the usefulness of our proposal relative to that benchmark. We also see that the "naive" uncorrected selective test generally has nontrivial size distortion, which confirms the necessity of adopting our novel critical value. In light of these findings, we recommend the selective test for practical applications.

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

4 An Empirical Application on Inflation Forecasting

4.1 The setting

We illustrate the proposed selective test in an empirical setting as described in Example 1, concerning the conditional evaluation of forecasting methods. By directly attacking the functional hypothesis (2.1), the test may be regarded as a formal nonparametric version of the "parametrized" test proposed in the original work of [27]. In this exercise, we take the user's desire for carrying out conditional evaluation as given and refer the reader to [27] for a discussion on the trade-off between the conditional and unconditional evaluation approaches.

We consider forecasts for the U.S. inflation measured by the monthly Consumer Price Index (CPI). Forecasting inflation is evidently of great interest to households, businesses, and policymakers. It is especially important for the conduct of money policy since the key role of many central banks is to maintain price stability. As noted by [40], accurately forecasting inflation has been a challenging task and some conventional econometric models have difficulty in beating even the simple random walk model. Meanwhile, machine learning methods combined with "big data" have been shown to deliver superior predictive ability in many applied areas. In a recent paper, [37] demonstrate the value of such methods for forecasting inflation. Set against this background, we compare the conditional performance of six forecasting methods: the random walk model (RW), the autoregressive (AR) model with the autoregressive order determined by the Bayesian information criterion, the diffusion indexes (DI) approach of [39], the Lasso method developed in [41], the elastic net (ElNet) proposed by [45], and the random forest (RF) method of [13]. The implementation details for these well-known methods are relegated to the supplemental appendix to save space.

Our goal is to demonstrate the empirical applicability of the selective test and further highlight its advantage over the benchmark non-selective approach in a concrete real data setting. To clarify, we do not attempt to rely on this small-scale study to promote any specific forecasting method. We simply consider the current setting as a representative example for many similar forecasting or prediction problems that have attracted much attention in the recent accounting, economics, and finance literature, in face of the fast development and adoption of machine learning methods.

We obtain data for the CPI and various predictor variables from the FRED-MD dataset, which is a leading macroeconomic database constructed by [36].[18] The dataset spans the period from January 1960 to December 2021 and contains 117 variables that are free from any missing value. All variables are updated on a monthly basis and are transformed to achieve stationarity. Using this dataset, we apply the aforementioned six methods to construct τ-month-ahead forecasts for the CPI, with $\tau = 1$ or 3, corresponding to monthly or quarterly forecasting horizons. The forecasting models are estimated under a 180-month rolling window scheme. The (pseudo) out-of-sample evaluation sample then spans the 1990–2021 period, containing $n = 384$ monthly observations in total.

To carry out the evaluation, we employ the absolute deviation loss function $L(f, f^*) = |f - f^*|$. Denote the sequence of forecasts generated by the jth method by $(F_{j,t+\tau})_{1\leq t\leq n}$. The ex post predictive loss is then given by $L(F_{j,t+\tau}, CPI_{t+\tau})$. Following [27], the conditional equal predictive ability (CEPA) hypothesis for the (j, k) pair of forecasts corresponds to (2.1) with

$$Y_t = L(F_{j,t+\tau}, CPI_{t+\tau}) - L(F_{k,t+\tau}, CPI_{t+\tau}).$$

We consider a bivariate conditioning variable $X_t = (Y_{t-\tau}, MU_t)$, where $Y_{t-\tau}$ is the lagged loss differential and MU_t is the macroeconomic uncertainty index developed in [30]. Note that the lagged loss differential was also used by [27], serving as the only conditioning variable in their empirical analysis. Here, we further include the macroeconomic uncertainty index to enrich the conditioning information set. This is an empirically relevant way to make the task of nonparametric conditional evaluation more challenging and allows us to better highlight the benefit of adopting the proposed selective test.

[18] The FRED-MD dataset is available from Michael McCraken's webpage; see https://research.stlouisfed.org/econ/mccracken/fred-databases/.

4.2 Empirical results

For each pair of competing forecasts and for each horizon τ, we test the CEPA hypothesis at 10% significance level by using both the selective test and the non-selective test. The six forecasting methods generate 15 pairwise comparisons in total. The tests are implemented following the same steps as in the simulation, except that we use the Newey–West estimator for the $\Sigma_{\mathcal{S}}$ matrix when $\tau = 3$, with the Newey–West bandwidth parameter chosen as the integer part of $1.2n^{1/3}$. In particular, the dictionary $\mathcal{M}$ of candidate bivariate approximating functions is formed using tensor products of univariate cubic spline functions up to p terms. In the empirical study, we consider $p \in \{4, \dots, 8\}$. The corresponding size of the dictionary ranges from $m = 15$ to 45.

Recall from the Monte Carlo study that the non-selective test can be very sensitive to the number of series terms m because it tends to reject more often when m is large even under the null hypothesis. Meanwhile, the selective test is notably more robust to the user-specified dictionary because it only relies on a much smaller subset of relevant approximating functions. The same phenomenon still obtains under the present real data setting, as we shall show below.

[Figure 2 Here]

To summarize how the selective and non-selective tests depend on the user-specified dictionary, we plot on Figure 2 the rejection rates (computed as the proportion of rejections across the 15 pairwise tests) as a function of the dictionary size m. From the figure, we observe that the non-selective test indeed rejects quite often when m is large, reflecting the severe size distortion of the non-selective test seen in the simulations. In particular, when $m = 45$, the non-selective test almost always rejects. This could be worrisome in empirical work because, without being aware of the size distortion issue, the applied researcher would have concluded that the test becomes more powerful when more approximating functions are used in the nonparametric inference. In view of the rejection pattern of the non-selective test, it also appears difficult to devise a reasonable rule-of-thumb that can satisfactorily guard against the over-rejection issue and mitigate its sensitivity with respect to the choice of the approximating functions.

In contrast, Figure 2 also shows that the rejection rate of the proposed selective test is quite stable across the board.[19] This feature is desirable in empirical work, especially when the empiricist wants to conduct a nonparametric test for a new empirical problem (e.g., evaluating some complicated nonlinear forecasting methods using a new dataset). In such a scenario, the empiricist is unlikely to have a good sense a priori about in which "direction" the functional null hypothesis may be violated, and so, needs to consider a large dictionary of candidate approximating functions. The selective test *allows* the empiricist to do so without running into the issues encountered by the non-selective test as discussed above.

The observed stability of the selective test not only holds on average (in terms of rejection rates as shown in Figure 2), but also obtains for individual tests (in terms of the rejection decisions). To see this, we report in Table 2 the p-value for each selective test, with rejections at the 10% significance level highlighted in bold. To save space, we focus on the cases with $p = 4$, 6, and 8, corresponding to $m = 15$, 28, and 45, respectively. We see that varying m leads to little changes in the rejection decision. This further highlights the robustness afforded by the selective test.

Finally, we comment on the relative performance of the six forecasting methods under evaluation. Rejecting the (two-sided) CEPA hypothesis signifies the difference in the state-dependent performances of the competing methods. To gain further information on the "direction" of the rejection, we follow the strategy of [27] and report in Table 2 the proportion of times that the method in the column outperforms (gauged by the estimated conditional expectation of loss differential) the method in the row. Based on the said testing results and auxiliary directional information, we highlight three findings. First, in all scenarios, the random walk forecast is significantly different from the others and exhibits clear underperformance. Second, the conventional AR and DI methods tend to underperform the other three machine-learning methods (i.e., Lasso, ElNet, and RF). Finally, it is difficult to statistically distinguish the conditional performances of Lasso, ElNet, and RF, although the RF method often attains the lowest conditional expected loss.

[19] The fact that the non-selective test rejects more often than the selective test does not mean that the former is more powerful. Rather, this is likely due to the size distortion of the non-selective test as we have seen in the simulation study.

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 [10] 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 $b_{\mathcal{S}}(\mathrm{s})$, $b'_{l,\mathcal{S}}(\mathrm{s})$, $b'_{u,\mathcal{S}}(\mathrm{s})$, $A_{\mathcal{S}}$, and $A'_{\mathcal{S}}$ that are needed in the definition of $\Pi(\mathcal{S}, \mathrm{s}, \lambda_n)$; (ii) an alternative choice of penalty parameters $\lambda_{\gamma,n}$ and $\{\omega_{\gamma,j}\}_{j\in\mathcal{M}_0^c}$ which satisfy Assumption 4; and (iii) theoretical justification of the recommended λ_n and $\{\omega_j\}_{j\in\mathcal{M}_0^c}$ in (2.9) and the alternative choice of penalty parameters $\lambda_{\gamma,n}$ and $\{\omega_{\gamma,j}\}_{j\in\mathcal{M}_0^c}$.

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 $\mathrm{s} \in \{\pm 1\}^{|\mathcal{S}\setminus\mathcal{M}_0|}$. These quantities are used to define the polytope $\Pi(\mathcal{S}, \mathrm{s}, \lambda_n)$ and its proxy $\widehat{\Pi}(\mathcal{S}, \mathrm{s}, \lambda_n)$. 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

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

Other choice of Lasso penalty parameters and justifications. The alternative choice of penalty parameters $\lambda_{\gamma,n}$ and $\{\omega_{\gamma,j}\}_{j\in\mathcal{M}_0^c}$, which satisfy Assumption 4, are presented in the algorithm below. This is followed by their theoretical justification, as well as the justification of the λ_n and $\{\omega_j\}_{j\in\mathcal{M}_0^c}$ in (2.9).

Algorithm A (Alternative Choice of Penalty Parameters)

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

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

Step 2. Set the penalty sequence in (2.8) as $\lambda_{\gamma,n} = \|\hat{\gamma}\|_0 \log(m) \log(\log(n))$.

Step 3. Set the weights in (2.8) as $\omega_{\gamma,j} = \hat{\sigma}_\gamma \|\mathbf{P}_j\|_n (\|\mathbf{P}_j\|_n |\hat{\gamma}_j / \hat{\sigma}_\gamma| + n^{-1/2})^{-1}$ for each $j \in \mathcal{M}_0^c$, where $\hat{\sigma}_\gamma$ denotes the sample standard deviation of $Y_t - P(X_t)^\top \hat{\gamma}$.

Proposition 1. *Suppose that: (i)* $||A'_{\mathcal{M}_0} \mathbf{P}^\top \boldsymbol{\epsilon}||_\infty = O_p((n \log(m))^{1/2})$*; (ii)* $n^{-1} \sum_{t=1}^n \epsilon_t^2 = \sigma_\epsilon^2 + o_p(1)$ *for some* σ_ϵ^2 *bounded from above and away from zero; (iii)* $|\mathcal{M}^\star| \log(m) \log(\log(n)) n^{-1} = o(1)$*. Then under Assumptions 1 and 2(i, ii), the penalty parameters* λ_n *and* $(\omega_j)_{j \in \mathcal{M}_0^c}$ *defined in (2.9) satisfy Assumptions 2(iii) and 3. Moreover if*

$$\frac{|\mathcal{M}^\star| \sqrt{\log(m) \log(\log(n))}}{n^{1/2} \min_{j \in \mathcal{R}} |b^*_{n,j}|} = o(1), \tag{A.2}$$

then the penalty parameters $\lambda_{\gamma,n}$ *and* $(\omega_{\gamma,j})_{j \in \mathcal{M}_0^c}$ *described in Algorithm A satisfy Assumptions 2(iii) and 4.*

References

[1] Donald W. K. Andrews. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica*, 59(2):307–345, 1991.

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

[3] Donald WK Andrews. Asymptotic optimality of generalized cl, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, 47(2-3):359–377, 1991.

[4] Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference after estimation of breaks. *Journal of Econometrics*, 224(1):39–59, 2021.

[5] Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. Technical report, National Bureau of Economic Research, 2021.

[6] Alexandre Belloni and Victor Chernozhukov. High dimensional sparse econometric models: An introduction. In *Inverse Problems and High-Dimensional Estimation*, pages 121–156. Springer, 2011.

[7] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Iván Fernández-Val. Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1):4–29, 2019.

[8] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345 – 366, 2015. High Dimensional Problems in Econometrics.

[9] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Ying Wei. Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *Annals of statistics*, 46(6B):3643, 2018.

[10] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

[11] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732, 2009.

[12] Zdravko I. Botev. The normal law under linear restrictions: Simulation and estimation via minimax tilting. *arXiv preprint arXiv:1603.04166*, 2016.

[13] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[14] Matias D Cattaneo, Max H Farrell, and Yingjie Feng. Large sample properties of partitioning-based series estimators. *The Annals of Statistics*, 48(3):1718–1741, 2020.

[15] Matias D Cattaneo, Michael Jansson, and Whitney K Newey. Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34(2):277–301, 2018.

[16] Matias D Cattaneo, Michael Jansson, and Whitney K Newey. Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361, 2018.

[17] Matias D Cattaneo, Ricardo P Masini, and William G Underwood. Yurinskii's coupling for martingales. *arXiv preprint arXiv:2210.00362*, 2022.

[18] Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. In J. James Heckman and E. Edward Leamer, editors, *Handbook of Econometrics*, volume 6B, chapter 76. Elsevier, 1 edition, 2007.

[19] Xiaohong Chen and Timothy M. Christensen. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465, 2015.

[20] Denis Chetverikov, Zhipeng Liao, and Victor Chernozhukov. On cross-validated lasso. *arXiv preprint arXiv:1605.02214*, 2016.

[21] Denis Chetverikov, Zhipeng Liao, and Victor Chernozhukov. On cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317, 2021.

[22] Denis Chetverikov and Jesper Riis-Vestergaard Sørensen. Selecting penalty parameters of high-dimensional m-estimators using bootstrapping after cross-validation. *arXiv preprint arXiv:2104.04716*, 2021.

[23] Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

[24] Charles Engel. The forward discount anomaly and the risk premium: A survey of recent evidence. *Journal of Empirical Finance*, 3(2):123–192, 1996.

[25] Charles Engel. Chapter 8 - exchange rates and interest parity. In Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, editors, *Handbook of International Economics*, volume 4, pages 453–522. Elsevier, Amsterdam, Netherlands, 2014.

[26] Eugene F. Fama. Forward and spot exchange rates. *Journal of Monetary Economics*, 14(3):319–338, 1984.

[27] Raffaella Giacomini and Halbert White. Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578, 2006.

[28] Lars Peter Hansen and Kenneth J. Singleton. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 50(5):1269–1286, 1982.

[29] Koen Jochmans. Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, 117(538):887–896, 2022.

[30] Kyle Jurado, Sydney C. Ludvigson, and Serena Ng. Measuring uncertainty. *American Economic Review*, 105(3):1177–1216, 2015.

[31] Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[32] Jia Li and Zhipeng Liao. Uniform nonparametric inference for time series. *Journal of Econometrics*, 219(1):28–51, 2020.

[33] Ker-Chau Li. Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, pages 1101–1112, 1986.

[34] Zhipeng Liao and Xiaoxia Shi. A nondegenerate vuong test and post selection confidence intervals for semi/nonparametric models. *Quantitative Economics*, 11(3):983–1017, 2020.

[35] Adam McCloskey. Hybrid confidence intervals for informative uniform asymptotic inference after model selection. *Biometrika*, 111(1):109–127, 2024.

[36] Michael W. McCracken and Serena Ng. FRED-MD: a monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

[37] Marcelo C. Medeiros, Gabriel F. R. Vasconcelos, Álvaro Veiga, and Eduardo Zilberman. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119, 2021.

[38] Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147 – 168, 1997.

[39] James H. Stock and Mark W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002.

[40] James H. Stock and Mark W. Watson. Modeling inflation after the crisis. Technical report, National Bureau of Economic Research, 2010.

[41] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[42] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[43] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

[44] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[45] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[46] Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751, 2009.

Tables and Figures

Table 1: Monte Carlo 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			
$m = 15$	0.056	0.052	0.052	0.051	0.054	0.047
$m = 28$	0.059	0.057	0.049	0.057	0.055	0.054
$m = 45$	0.067	0.058	0.049	0.064	0.059	0.050
			Panel B: Non-Selective Test			
$m = 15$	0.194	0.118	0.067	0.190	0.110	0.068
$m = 28$	0.567	0.327	0.159	0.577	0.343	0.166
$m = 45$	0.688	0.670	0.450	0.722	0.681	0.446
			Panel C: Uncorrected (Selective) Test			
$m = 15$	0.069	0.062	0.056	0.066	0.060	0.054
$m = 28$	0.099	0.084	0.062	0.092	0.080	0.073
$m = 45$	0.124	0.095	0.075	0.115	0.102	0.080

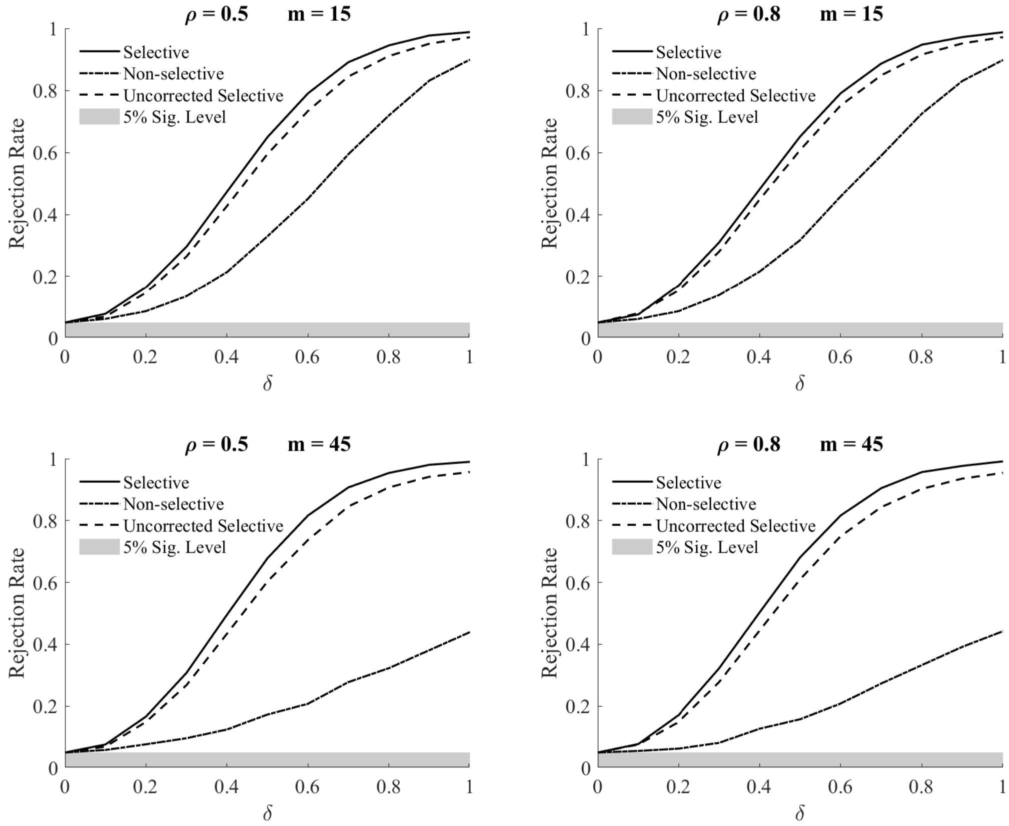
Note: This table reports the rejection rates of the selective test, the non-selective test, and the uncorrected selective test at the 5% significance level under the null hypothesis (i.e., $\delta = 0$). These results are generated for a variety of specifications under which the autoregressive coefficient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

Table 2: Selective Tests of Conditional Equal Predictive Ability for Inflation Forecasts

	$m = 15$					$m = 28$					$m = 45$				
	RW	AR	DI	Lasso	ElNet	RW	AR	DI	Lasso	ElNet	RW	AR	DI	Lasso	ElNet
Benchmark						*Panel A: one-month-ahead forecasts*									
AR	**0.000** **(0.00)**					**0.000** **(0.00)**					**0.003** **(0.00)**				
DI	**0.001** **(0.00)**	0.925 (1.00)				**0.000** **(0.00)**	0.924 (1.00)				**0.001** **(0.00)**	0.922 (1.00)			
Lasso	**0.000** **(0.00)**	**0.048** **(0.00)**	**0.095** **(0.00)**			**0.000** **(0.00)**	**0.048** **(0.00)**	**0.089** **(0.00)**			**0.000** **(0.00)**	**0.049** **(0.00)**	**0.096** **(0.00)**		
ElNet	**0.000** **(0.00)**	**0.034** **(0.00)**	**0.086** **(0.00)**	0.871 (0.81)		**0.000** **(0.00)**	**0.038** **(0.00)**	**0.089** **(0.00)**	0.743 (0.00)		**0.000** **(0.00)**	**0.032** **(0.00)**	**0.086** **(0.00)**	0.745 (0.00)	
RF	**0.000** **(0.00)**	**0.006** **(0.00)**	**0.049** **(0.00)**	0.491 (0.29)	0.643 (0.00)	**0.000** **(0.00)**	**0.008** **(0.00)**	**0.049** **(0.00)**	0.716 (0.29)	0.651 (0.00)	**0.000** **(0.00)**	**0.005** **(0.00)**	**0.054** **(0.00)**	0.593 (0.00)	0.648 (0.00)
						Panel B: three-month-ahead forecasts									
AR	**0.000** **(0.00)**					**0.000** **(0.00)**					**0.000** **(0.00)**				
DI	**0.001** **(0.00)**	0.157 (1.00)				**0.005** **(0.00)**	0.163 (1.00)				**0.001** **(0.00)**	0.170 (1.00)			
Lasso	**0.000** **(0.00)**	0.290 (0.00)	**0.056** **(0.00)**			**0.000** **(0.00)**	0.294 (0.00)	**0.063** **(0.00)**			**0.000** **(0.00)**	0.295 (0.00)	**0.066** **(0.00)**		
ElNet	**0.000** **(0.00)**	0.268 (0.00)	**0.053** **(0.00)**	0.605 (0.00)		**0.000** **(0.00)**	0.275 (0.00)	**0.060** **(0.00)**	0.615 (0.00)		**0.000** **(0.00)**	0.278 (0.00)	**0.063** **(0.00)**	0.616 (0.00)	
RF	**0.000** **(0.00)**	0.261 (0.00)	**0.035** **(0.00)**	0.693 (0.00)	0.724 (0.00)	**0.000** **(0.00)**	0.256 (0.00)	**0.042** **(0.00)**	0.694 (0.00)	0.732 (0.00)	**0.000** **(0.00)**	0.264 (0.00)	**0.042** **(0.00)**	0.684 (0.00)	0.725 (0.00)

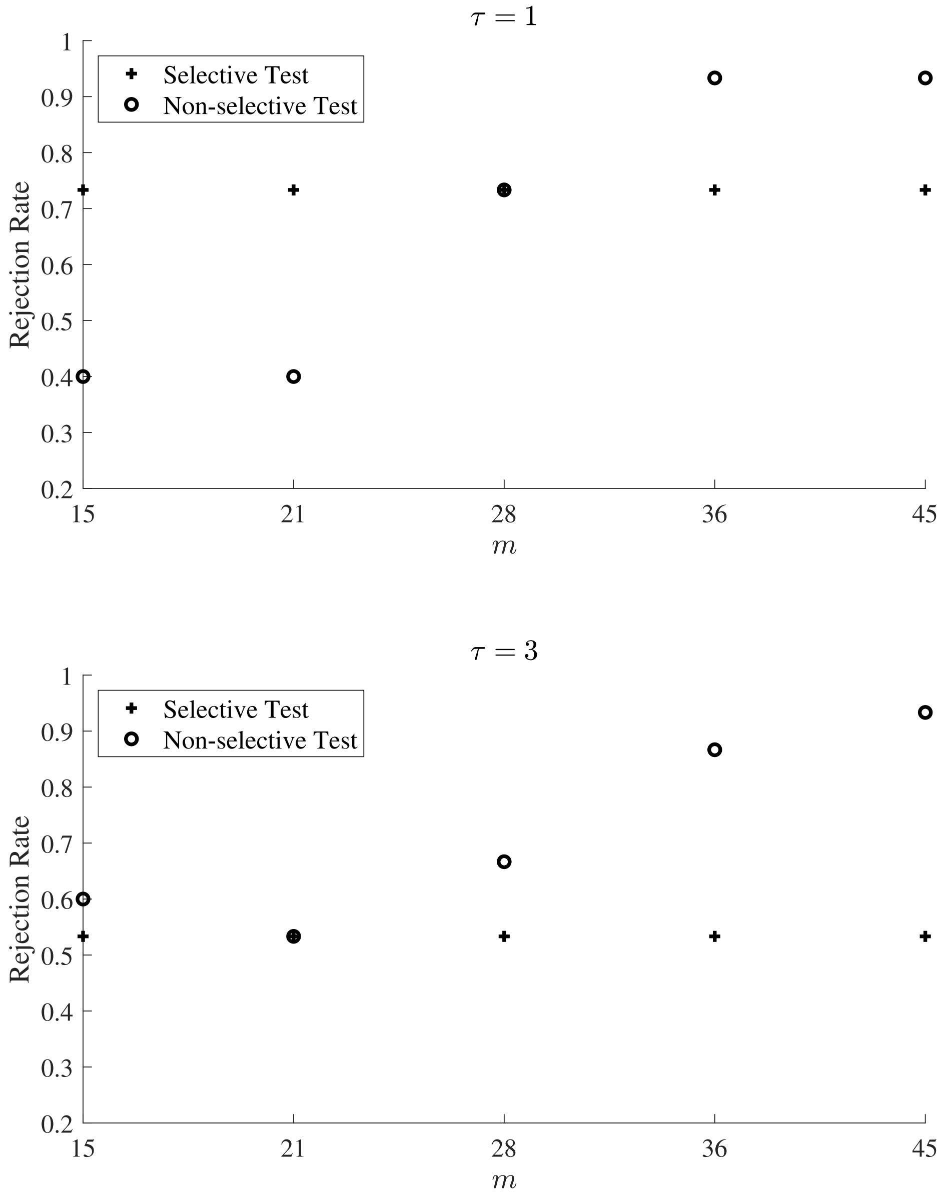
Note: This table reports results for the pairwise selective tests among six inflation forecasting methods. The main entries are p-values of the test with respect to the null hypothesis that the methods in each row and column have equal conditional predictive ability under the absolute deviation loss. Rejections at 10% significance level are highlighted in bold. The numbers in parentheses indicate the proportions of times that the conditional expected loss of the method in the column is lower than that of the method in the row.

Figure 1: Simulation Results: Size-adjusted Power Curves



Note: This figure plots the size-adjusted Monte Carlo rejection rates of the selective test (solid), the non-selective test (dotted), and the uncorrected selective test (dashed) at the 5% significance level (highlighted by the shaded area) over $\delta \in \{0, 0.1, 0.2, \ldots, 1\}$. Results for $m = 15$ (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.

Figure 2: Rejection Rates of Selective and Non-Selective Tests for Inflation Forecasts



Note: This figure reports the rejection rates of 10% level selective and non-selective tests (averaged across the 15 pairwise comparisons among six forecasting methods) as functions of the number of candidate approximating functions m in the dictionary.