Conditional Superior Predictive Ability*

Jia Li†
Duke University and Singapore Management University

Zhipeng Liao‡
UCLA

Rogier Quaedvlieg§
Erasmus School of Economics

June 10, 2021

Abstract

This paper proposes a test for the conditional superior predictive ability (CSPA) of a family of forecasting methods with respect to a benchmark. The test is functional in nature: Under the null hypothesis, the benchmark's conditional expected loss is no more than those of the competitors, uniformly across all conditioning states. By inverting the CSPA tests for a set of benchmarks, we obtain confidence sets for the uniformly most superior method. The econometric inference pertains to testing conditional moment inequalities for time series data with general serial dependence, and we justify its asymptotic validity using a uniform nonparametric inference method based on a new strong approximation theory for mixingales. The usefulness of the method is demonstrated in empirical applications on volatility and inflation forecasting.

Keywords: conditional moment inequality, forecast evaluation, inflation, intersection bounds, machine learning, volatility.

*We thank the Co-Editor (Francesca Molinari) and four anonymous referees for their comments and suggestions, which have greatly improved the paper. We also thank Raffaella Giacomini, Jinyong Hahn, Peter Reinhard Hansen (discussant), Hyungsik Roger Moon and conference and seminar participants at Aarhus, CREST, the 2018 Triangle Econometrics Conference, Southern California Winter Econometrics Day, the 2019 Toulouse Financial Econometrics Conference, and the 2021 SoFiE seminar for their comments. Liao's research was partially supported by National Science Foundation Grant SES-1628889. Quaedvlieg was financially supported by the Netherlands Organisation for Scientific Research (NWO) Grant 451-17-009.

†Corresponding author: 90 Stamford Rd, Singapore 178903; e-mail: jiali@smu.edu.sg

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

§P.O. Box 1738, 3000DR Rotterdam, Netherlands; e-mail: quaedvlieg@ese.eur.nl.

JEL Code: C14, C22.

1 Introduction

A central problem in time-series econometrics is forecasting economic quantities, such as GDP growth, inflation, stock returns, and volatility. Empiricists often face an extensive list of "reasonable" candidate forecasting methods that are invariably backed by influential prior studies. For example, classical ARMA- and ARCH-type models contain numerous variants, and the recent trend of using machine-learning algorithms—such as LASSO, random forests, support vector machines, and deep neural nets—can make a forecaster's choice even more difficult. Rigorously evaluating the relative performance of these methods, and identifying superior ones, is thus of great and ever-growing importance.

The most popular forecast evaluation method is, arguably, Diebold and Mariano's (1995) test. Under the null hypothesis of the Diebold–Mariano test, two competing methods have the same unconditional expected loss, and the test can be carried out using a simple t-test. More generally, a multivariate extension of the Diebold–Mariano test can be used to test *unconditional equal predictive ability (UEPA)* concerning multiple competing forecasts, which amounts to testing a system of unconditional moments (i.e., expected loss differentials) being zero.

Two significant extensions of the Diebold–Mariano test have been developed in the literature. The first is the test for *unconditional superior predictive ability (USPA)*, which is first studied by White (2000) and later refined by Hansen (2005). The null hypothesis states that a benchmark method weakly outperforms a collection of competing alternatives, as formalized by a system of unconditional moment inequalities. In his seminal work, White (2000) proposes critical values under the least favorable null (i.e., all inequalities are binding), which effectively reduces the USPA test into a multivariate version of the Diebold–Mariano test. Hansen makes an important observation that White's test can be very conservative when there are competing methods that are clearly inferior to the benchmark. To remedy this issue, Hansen proposes a preliminary selection based on studentized moment conditions to remove these clearly inferior methods, and shows that this can significantly improve the test's statistical power.

The second extension is the *conditional equal predictive ability (CEPA)* test proposed by Giacomini and White (2006). The authors eloquently argue that, in practice, a forecast evaluator is not only interested in knowing whether one method is better than another on average, but also

interested in when this occurs. This consideration is particularly relevant when the methods on the forecaster's "shortlist" appear similar on average, but can actually behave very differently conditional on certain economic states.[1] The CEPA null hypothesis states that conditional expected loss functions of different forecasting methods are identically the same across all conditioning states. The corresponding econometric inference in principle concerns global features of the conditional expected loss functions. Giacomini and White, however, do not directly attack this functional inference problem. Instead, they propose a practical method based on a fixed number of (instrumented) unconditional moments implied by the original conditional ones. As such, the Giacomini–White test is operationally the same as a finite-dimensional UEPA test.

Set against this background, we extend these existing evaluation paradigms by proposing a test for *conditional superior predictive ability (CSPA)*, which synergizes the key insights of White (2000), Hansen (2005), and Giacomini and White (2006). Specifically, the CSPA null hypothesis asserts the uniform (weak) superiority of the benchmark method, in the sense that the conditional expected loss of the benchmark does not exceed those of the competing forecasts across all conditioning states. On the other hand, a rejection of the CSPA null hypothesis suggests that some competing alternative method outperforms the benchmark in certain states, which are revealed as a by-product of the testing procedure, providing useful diagnostic information.

Our CSPA test formally pertains to testing inequalities for forecasting methods' conditional expected loss functions. To implement the test, we nonparametrically estimate the conditional mean function using the series method (Andrews (1991a), Newey (1997), Chen (2007)), and then adopt the intersection-bound methodology (Chernozhukov, Lee, and Rosen (2013)) to conduct inference regarding the functional inequalities. It is well known that the underlying uniform inference problem is non-Donsker, for which conventional functional central limit theorems based on the weak convergence concept are not applicable. In a setting with independent data, Chernozhukov, Lee, and Rosen (2013) address this issue by using Yurinskii's coupling, which provides a strong Gaussian approximation for the growing-dimensional moment conditions in the series estimation. As a result, the t-statistic process (indexed by the conditioning state variable) can be strongly approximated by a divergent Gaussian process that can be used to construct critical values.

In this paper, we also adopt the strong approximation strategy to make inference on the conditional moment inequalities, but in a more general time-series setting. Specifically, we allow the data to be generated as mixingales, which include martingale differences, linear processes, and various types of mixing and near-epoch dependent processes as special cases, and naturally accom-

[1] A related issue is the evaluation of dynamic stochastic general equilibrium model's predictive density forecasts; see, for example, Herbst and Schorfheide (2012).

modate data heterogeneity (see Davidson (1994)). One possible way to achieve this is to invoke the strong approximation theory recently developed by Li and Liao (2020), which provides a Yurinskii coupling for general dependent data modeled as mixingales, and then proceed as in Chernozhukov, Lee, and Rosen (2013). However, a drawback of this approach is that Yurinskii's coupling concerns the approximation for the entire sample moment vector (under Euclidean distance), and hence, occurs at a slow rate of convergence. This in turn leads to relatively restrictive conditions on how fast the number of approximating series terms, m_n, can grow with the sample size n. Specifically, in both the independent-data setting of Chernozhukov, Lee, and Rosen (2013) and the time-series setting of Li and Liao (2020), Yurinskii's coupling is available when $m_n = o(n^{1/5})$. This issue can be more severe in the time-series context (which is also the setting here), as the requisite restriction on m_n becomes more stringent when the data is serially "more dependent."

Motivated by this issue, in this paper we develop a new strong approximation theory in the time-series setting which substantially improves the theory of Li and Liao (2020) for making uniform series inference in the time-series context. Rather than strongly approximating the entire vector of moment conditions, we establish a strong approximation for the "maximum projection" of this growing-dimensional statistic along a large number of directions. This new result is weaker than Yurinskii's coupling, and it can be established under notably weaker restrictions on the number of series terms. The "cost" of not coupling the entire vector of sample moments is that one can no longer construct a strong approximation for the t-statistic process associated with the (functional) series estimator, which is needed as a high-level condition in Chernozhukov, Lee, and Rosen's (2013) theory (see Condition C.2 of that paper). That being said, we show that the general framework of intersection-bound inference can nevertheless be adapted to accommodate this weaker notion of coupling. Our theory thus extends that of Chernozhukov, Lee, and Rosen (2013) by both allowing for time-series dependence in the data and a weaker notion of strong approximation. This theory may be further extended to settings with spatial dependence, but that extension is beyond the scope of the present paper on forecast evaluation.[2]

These new econometric ingredients suggest that the proposed (nonparametric) CSPA test differs from the conventional forecast evaluation methods not only in concept, but also in econometric technicality: The unconditional tests of Diebold and Mariano (1995), White (2000), and Hansen (2005) concern a fixed number of unconditional expected losses; Giacomini and White's (2006) CEPA hypothesis is nonparametric in nature, but they only test a fixed number of implied unconditional moments. In contrast, the CSPA test directly addresses the functional inference by

[2]As another extension, one may also use a kernel estimator as an alternative to the series estimator, provided that a similar strong approximation theory can be established for the former in a general time-series setting.

adopting and generalizing recent results from the partial identification literature (see Molinari (2019) for a recent review).[3] CSPA unifies these prior evaluation paradigms in the following theoretical sense: USPA is its special case with empty conditioning information set, and CEPA corresponds to the least favorable null hypothesis of CSPA. It is interesting to note that, like Hansen's (2005) USPA test, our CSPA test also involves a preliminary selection that "effectively removes" clearly inferior forecasts in the computation of critical values and, in the same spirit, it also removes regions in the state space on which the competing forecasts are clearly inferior, which is unique to our functional inference problem.

We apply the CSPA test in two important empirical settings to demonstrate its usefulness. The first pertains to the evaluation of volatility forecasts, which is one of the most important topics in financial econometrics. We consider a variety of autoregressive (AR) models for realized volatility, including: AR(1), AR(22) with or without adaptive variable selection, fractionally integrated AR, HAR (Corsi (2009)) and its HARQ extension (Bollerslev, Patton, and Quaedvlieg (2016, 2018)). Consistent with prior studies, we find that HARQ is generally superior to the other methods using Hansen's (2005) USPA test. But the CSPA test provides further useful diagnostic information. We find that in many cases, we cannot reject the CSPA null hypothesis that HARQ weakly dominates the other methods uniformly across different states. Interestingly, the conditional test also reveals cases in which alternative methods—particularly the fractionally integrated model and, somewhat surprisingly, the simple AR(1) model—significantly outperforms HARQ over certain regions of the state space. With the CSPA criterion, we pose a new challenge for the empirical search of "uniformly" superior volatility forecasting methods, for which the proposed test can be used to formally run the horse race.

In the second empirical application, we evaluate inflation forecasts in a macroeconometric setting. We consider eight forecasting methods for monthly inflation based on the recent work of Medeiros, Vasconcelos, Veiga, and Zilberman (2019). Four of these methods are traditional inflation models (e.g., factor model and Bayesian vector autoregression). The other four are machine-learning algorithms. Using Hansen's (2005) USPA test, we find that the traditional models are typically beaten by at least one of the four machine-learning methods, but the latter methods appear to be virtually indistinguishable judged by (unconditional) average loss. In contrast, the CSPA test helps distinguishing them for a wide variety of economically important conditioning

[3]Our theory is a complement, instead of substitute, of Giacomini and White's (2006) test. In fact, the strong approximation theory developed here can be used to provide a nonparametric interpretation of Giacomini and White's test, by allowing the number of instruments to grow with the sample size. We do not develop this explicitly here so as to be focused on the CSPA test.

variables.

The paper is organized as follows. Section 2 describes our test and asymptotic theory. Section 3 reports simulation results. Sections 4 and 5 present the empirical applications on volatility and inflation forecasting, respectively. Section 6 concludes. The appendix contains proofs for our main theoretical results. Additional theoretical and numerical results are detailed in the Supplemental Appendix to this paper.

2 Testing for conditional superior predictive ability

We present the theory for the CSPA test in this section. Section 2.1 introduces the hypotheses of interest. In Section 2.2, we describe the CSPA test and establish its asymptotic validity under high-level conditions. Section 2.3 further shows how to invert the CSPA test to obtain confidence sets for the most superior forecasting method. Our econometric inference relies on a new coupling theory established in Section 2.4, which may be skipped by readers who are mainly interested in applications. Below, for two real sequences a_n and b_n, we write $a_n \asymp b_n$ if $b_n/C \leq a_n \leq Cb_n$ for some constant $C > 1$.

2.1 Forecast evaluation hypotheses

Let $(F_t^\dagger)_{t\geq 1}$ be the time series to be forecast. We consider a benchmark forecast sequence $(F_{0,t})_{1\leq t\leq n}$ and J competing forecast sequences $(F_{j,t})_{1\leq t\leq n}$, $1 \leq j \leq J$. With a loss function $L(\cdot,\cdot)$, we measure the performance of the benchmark method relative to the jth competing alternative by the loss differential sequence defined as

$$Y_{j,t} \equiv L\left(F_t^\dagger, F_{j,t}\right) - L\left(F_t^\dagger, F_{0,t}\right). \tag{2.1}$$

In particular, $Y_{j,t} \geq 0$ indicates that the benchmark (weakly) outperforms method j in period t.[4] We stress from the outset that, in this paper we focus on the evaluation of forecasting methods by taking the forecasts and their loss differentials as primitives. This is in the same spirit as Diebold and Mariano (1995) and Giacomini and White (2006), and is distinct from the evaluation of forecasting models (West (1996)); see Diebold (2015) for a comprehensive discussion on the distinction between these two evaluation problems.

[4]The choice of loss function depends on the user's own assessment of the "cost" of a forecasting error. It is thus possible that the user's loss function is different from the forecaster's. In this case, an "optimal" forecast under one loss function may be outperformed by the other forecast if the comparison is carried out using a different loss function. For example, the conditional mean is optimal under the quadratic loss, but is generally inferior to the conditional median under the absolute deviation loss.

Two types of null hypotheses are commonly tested in forecast evaluation. One is the hypothesis of unconditional equal predictive ability (UEPA):

$$H_0^{UEPA} : \mathbb{E}\left[Y_{j,t}\right] = 0, \quad 1 \leq j \leq J, \tag{2.2}$$

that is, the benchmark has the same expected performance as all competing alternatives. The other is the hypothesis of unconditional superior predictive ability (USPA):

$$H_0^{USPA} : \mathbb{E}\left[Y_{j,t}\right] \geq 0, \quad 1 \leq j \leq J, \tag{2.3}$$

meaning that the benchmark weakly outperforms the others. UEPA and USPA are clearly connected, in that the former is the least-favorable null of the latter.

The unconditional tests are informative about the relative performance of forecasting methods on average. As such, they have a "blind spot:" Two methods may appear to have (statistically) identical performance on average, but can behave very differently given certain economic conditions. Giacomini and White (2006) advocate testing the hypothesis of conditional equal predictive ability (CEPA), that is,

$$H_0^{CEPA} : \mathbb{E}\left[Y_{j,t} | X_t = x\right] = 0, \quad x \in \mathcal{X}, \quad 1 \leq j \leq J, \tag{2.4}$$

where X_t is a conditioning state variable chosen by the evaluator, and $\mathcal{X}$ specifies the conditioning region as a subset of the domain of X. For example, one can track forecasting methods' performance through business cycles by setting X_t to be a cyclical indicator (e.g., GDP growth). The CEPA null hypothesis then states that the benchmark performs equally well as all competing alternatives, not only on average, but also through the ups-and-downs of the economy. This hypothesis is rejected if some competing forecasting method performs differently than the benchmark in some states (say, expansion or recession).

A rejection of the CEPA hypothesis is not directly informative about whether the competing methods is better or worse than the benchmark—it only signifies their difference. In contrast, we consider the *conditional superior predictive ability (CSPA)* hypothesis. The null hypothesis asserts

$$H_0^{CSPA} : \mathbb{E}\left[Y_{j,t} | X_t = x\right] \geq 0, \quad x \in \mathcal{X}, \quad 1 \leq j \leq J. \tag{2.5}$$

This imposes a very stringent requirement on the benchmark, that is, it needs to weakly dominate all competing methods across the conditioning region $\mathcal{X}$. Therefore, "passing" the CSPA test should be deemed a highly desirable feature of a forecasting method.[5]

[5]While Giacomini and White (2006) focus on the state-dependent performance of competing forecasts, another

Introducing CSPA to the forecast evaluation literature seems to be rather natural and conceptually straightforward: CSPA is to CEPA simply as USPA is to UEPA. However, testing this new hypothesis is fundamentally different from—and econometrically much more complicated than—those in the prior forecast evaluation literature. To see why, note that unconditional tests only concern the finite-dimensional vector $(\mathbb{E}[Y_{j,t}])_{1\leq j\leq J}$. In contrast, conditional tests involve functional inference for conditional expectation functions of loss differentials given by[6]

$$h_j(x) \equiv \mathbb{E}[Y_{j,t}|X_t = x], \quad 1 \leq j \leq J.$$

The related functional inference is theoretically nontrivial because it requires knowledge about the global behavior of the h_j functions. In their pioneering work, however, Giacomini and White (2006) bypassed the functional inference by instead testing certain implications of CEPA. These authors take as given a finite-dimensional instrument W_t that is measurable with respect to the σ-field generated by X_t, and derive from (2.4) the following unconditional moment equalities:

$$\mathbb{E}[Y_{j,t}W_t] = 0, \quad 1 \leq j \leq J.$$

These can then be tested by using a conventional Wald test.

Unlike Giacomini and White (2006), we attack the functional inference problem directly in our study of CSPA. Our approach relies on inference methods recently developed in the partial identification literature, particularly those concerning conditional moment inequalities. We adopt the intersection-bound approach originally proposed by Chernozhukov, Lee, and Rosen (2013) for microeconometric applications, and extend it to a general time-series setting. More precisely, we rewrite the CSPA hypothesis as

$$H_0^{CSPA} : \eta^* \equiv \min_{1\leq j\leq J} \inf_{x\in\mathcal{X}} h_j(x) \geq 0, \tag{2.6}$$

important strand of literature studies the forecasts' time-varying relative performance; see, for example, Giacomini and Rossi (2010, 2016), Rossi and Sekhposyan (2010), and Rossi (2013). Specifically, Giacomini and Rossi (2010) propose tests for the null hypothesis that two competing forecasts have the same expected loss at all times. Their test statistic is constructed using the moving average process of the loss differential series, instead of the full-sample average, and is able to detect certain types of temporal instability in the forecasts' relative performance. It is interesting to note that the moving average process resembles a nonparametric estimator of the time-varying mean, although Giacomini and Rossi (2010) do not consider a nonparametric theory. The strong approximation theory developed in the present paper may be adapted to extend Giacomini and Rossi's (2010) theory to a fully nonparametric setting.

[6]The conditional mean function $h_j(\cdot)$ is assumed to be time-invariant, that is, Y_t is conditionally mean-stationary given X_t. Note that this condition does not rule out nonstationarity in the data. For example, both X_t and Y_t processes may have time-varying unconditional means, and their higher moments are not restricted. Allowing $h_j(\cdot)$ to be time-varying might be an interesting topic for future research.

where η^* measures the worst relative performance of the benchmark with respect to all competitors over all conditioning states. This equivalent form of the CSPA null hypothesis highlights the fact that the CSPA inference concentrates on the infimum η^*. For some significance level $\alpha \in (0, 1/2)$, we shall construct a $1-\alpha$ upper confidence bound $\hat{\eta}_{n,1-\alpha}$ for η^* such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\eta^* \leq \hat{\eta}_{n,1-\alpha}\right) \geq 1-\alpha. \tag{2.7}$$

This implies that under H_0^{CSPA}, $\hat{\eta}_{n,1-\alpha} \geq \eta^* \geq 0$ holds with at least $1-\alpha$ probability asymptotically. Equivalently, a test that rejects H_0^{CSPA} if and only if $\hat{\eta}_{n,1-\alpha} < 0$ has probability of type-I error bounded by α in large samples. The feasible computation of $\hat{\eta}_{n,1-\alpha}$ and the theoretical properties of the test are detailed in Section 2.2 below.

2.2 The CSPA test and its asymptotic properties

In this subsection, we describe how to implement the CSPA test and establish its theoretical validity (see Proposition 1). For readability, we present the theory under high-level conditions, most of which are standard for series-based nonparametric inference and are well understood in the literature. The key exception is a condition for coupling a maximum projection of a growing-dimensional sample moment constructed using dependent data, for which we develop a new theory in Section 2.4.

To perform the CSPA test, we first estimate the $h_j(\cdot)$ functions nonparametrically by the least-square series regression.[7] Let $P(x) = (p_1(x), \ldots, p_{m_n}(x))^\top$ be an m_n-dimensional vector of approximating basis functions, such as polynomial, Fourier, spline, and wavelet series; see Chen (2007) for a comprehensive review. By convention, we suppose that $P(\cdot)$ always contains the constant function by setting $p_1(\cdot) \equiv 1$. To conduct series estimation, for each j, we regress $Y_{j,t}$ on $P(X_t)$ and obtain the regression coefficient

$$\hat{b}_{j,n} \equiv \widehat{Q}_n^{-1}\left(n^{-1}\sum_{t=1}^{n} P(X_t)\, Y_{j,t}\right), \quad \text{where} \quad \widehat{Q}_n \equiv n^{-1}\sum_{t=1}^{n} P(X_t)\, P(X_t)^\top.$$

The functional estimator for $h_j(\cdot)$ is then given by

$$\hat{h}_{j,n}(\cdot) \equiv P(\cdot)^\top \hat{b}_{j,n}.$$

The series regression coefficient $\hat{b}_{j,n}$ formally resembles the conventional least-square estimator, but corresponds to a very different theory. We require the number of series terms $m_n \to \infty$

[7] Our approach is inspired by Diebold and Mariano's (1995) suggestion that one may use regressions to examine whether the loss differential may be explained by other variables.

asymptotically so that the unknown $h_j(\cdot)$ function can be approximated sufficiently well by a large number of approximating functions. The growing dimension of $\hat{b}_{j,n}$ renders the conventional weak-convergence-based characterization of asymptotic normality inappropriate. This is why we shall need a strong approximation theory for growing-dimensional statistics to construct uniform inference, as in Chernozhukov, Lee, and Rosen (2013), Belloni, Chernozhukov, Chetverikov, and Kato (2015), and Li and Liao (2020).

We now proceed to describe the inference procedure. Let $u_t \equiv (u_{1,t}, \ldots, u_{J,t})^\top$, where $u_{j,t}$ is the nonparametric regression error term defined as

$$u_{j,t} \equiv Y_{j,t} - h_j(X_t).$$

We further set A_n to be the $Jm_n \times Jm_n$ covariance matrix of $n^{-1/2}\sum_{t=1}^n u_t \otimes P(X_t)$, that is,

$$A_n \equiv Var\left(n^{-1/2}\sum_{t=1}^n u_t \otimes P(X_t)\right), \tag{2.8}$$

where $\otimes$ denotes the Kronecker product. To conduct feasible inference, we suppose that a heteroskedasticity and autocorrelation consistent (HAC) estimator $\widehat{A}_n$ for A_n is available and satisfies the following condition, where $\|\cdot\|_S$ denotes the matrix spectral norm.

Assumption 1. *$\|\widehat{A}_n - A_n\|_S = O_p(\delta_{A,n})$ for some sequence $\delta_{A,n} \to 0$ at polynomial rate.*[8]

Assumption 1 is high-level and in fact nonstandard, because it concerns the convergence rate of HAC estimators with growing dimensions (i.e., $Jm_n \to \infty$), whereas the classical HAC estimation theory (e.g., Newey and West (1987) and Andrews (1991b)) is developed in settings with fixed dimensions. In the present growing-dimensional setting, the consistency of the HAC estimator is not enough for feasible inference, and we need a stronger form of consistency (i.e., with polynomial rate) as stated by the assumption above.

A theoretically valid choice of $\widehat{A}_n$ that verifies Assumption 1 is the Newey–West type HAC estimator (see Theorem 6 of Li and Liao (2020)). However, Newey–West estimators may lead to nontrivial size distortions in finite samples. This is well-known in the HAC estimation literature, and we also document a similar issue in our Monte Carlo experiments. To remedy this finite-sample distortion, in this paper we analyze a more general class of pre-whitened HAC estimators in the spirit of Andrews and Monahan (1992), and characterize their convergence rates in the growing-dimensional setting. We find that the pre-whitened estimator indeed leads to better size control. The theory on the pre-whitened estimator, however, is somewhat tangential to our main result on

[8] That is, $\delta_{A,n} \asymp n^{-a}$ for some fixed constant $a > 0$ that may be arbitrarily small.

CSPA testing, and it also requires a technical setup that is not used elsewhere in the paper. To remain focused, we relegate all details on the pre-whitened estimator to Supplemental Appendix SC.

Equipped with the estimator $\widehat{A}_n$, we can estimate the $Jm_n \times Jm_n$ covariance matrix of the normalized estimators $(n^{1/2}(\hat{b}_{j,n} - b^*_{j,n}))_{1\leq j\leq J}$ via

$$\widehat{\Omega}_n \equiv \left(I_J \otimes \widehat{Q}_n\right)^{-1} \widehat{A}_n \left(I_J \otimes \widehat{Q}_n\right)^{-1},$$

where I_J denotes the $J \times J$ identity matrix, and $b^*_{j,n}$ is the "population analogue" of $\hat{b}_{j,n}$ that is formally introduced in Assumption 2 below. We further partition $\widehat{\Omega}_n$ into $J \times J$ blocks of $m_n \times m_n$ submatrices $\widehat{\Omega}_n(j,k)$, $1 \leq j,k \leq J$. Note that $\widehat{\Omega}_n(j,k)$ is the estimator of the covariance matrix between $n^{1/2}(\hat{b}_{j,n} - b^*_{j,n})$ and $n^{1/2}(\hat{b}_{k,n} - b^*_{k,n})$. The standard deviation function of $n^{1/2}(\hat{h}_{j,n}(x) - h_j(x))$ is then estimated by

$$\hat{\sigma}_{j,n}(x) \equiv (P(x)^\top \widehat{\Omega}_n(j,j) P(x))^{1/2}.$$

Algorithm 1, below, provides the implementation details of the CSPA test, which is based on the intersection-bound inference of Chernozhukov, Lee, and Rosen (2013).

Algorithm 1 (Implementation of the CSPA Test).
Step 1. Simulate a Jm_n-dimensional random vector $\left(\xi_1^{*\top}, \ldots, \xi_J^{*\top}\right)^\top \sim \mathcal{N}(0, \widehat{\Omega}_n)$, where each ξ_j^* is m_n-dimensional. Set $\hat{t}^*_{j,n}(x) \equiv P(x)^\top \xi_j^* / \hat{\sigma}_{j,n}(x)$.
Step 2. Repeat step 1 many times. Set $\tilde{\gamma}_n \equiv 1 - 0.1/\log(n)$. Let $\hat{K}_n$ be the $\tilde{\gamma}_n$-quantile of $\max_{1\leq j\leq J} \sup_{x\in\mathcal{X}} \hat{t}^*_{j,n}(x)$ in the simulated sample and then set

$$\widehat{\mathcal{V}}_n \equiv \left\{(j,x) : \hat{h}_{j,n}(x) \leq \min_{1\leq j\leq J} \inf_{x\in\mathcal{X}} \left(\hat{h}_{j,n}(x) + n^{-1/2}\hat{K}_n\hat{\sigma}_{j,n}(x)\right) + 2n^{-1/2}\hat{K}_n\hat{\sigma}_{j,n}(x)\right\}.$$

Step 3. Set $\hat{k}_{n,1-\alpha}$ as the $(1-\alpha)$-quantile of $\sup_{(j,x)\in\widehat{\mathcal{V}}_n} \hat{t}^*_{j,n}(x)$ and set

$$\hat{\eta}_{n,1-\alpha} = \min_{1\leq j\leq J} \inf_{x\in\mathcal{X}} \left[\hat{h}_{j,n}(x) + n^{-1/2}\hat{k}_{n,1-\alpha}\hat{\sigma}_{j,n}(x)\right].$$

Reject the CSPA null hypothesis at significance level α if $\hat{\eta}_{n,1-\alpha} < 0$. □

It is instructive to provide some intuition for this procedure. In step 1, the Gaussian variables $(\xi_j^*)_{1\leq j\leq J}$ are generated to approximate the distribution of $(n^{1/2}(\hat{b}_{j,n} - b^*_{j,n}))_{1\leq j\leq J}$, and the (conditionally) Gaussian processes $(\hat{t}^*_{j,n}(\cdot))_{1\leq j\leq J}$ mimic the law of the t-statistic processes associated with the functional estimators $(\hat{h}_{j,n}(\cdot))_{1\leq j\leq J}$. Step 2 of the algorithm implements the adaptive inequality selection: We jointly select $j \in \{1,\ldots,J\}$ and $x \in \mathcal{X}$ into the set $\widehat{\mathcal{V}}_n$ such that, with

probability approaching one, $h_j(x)$ is minimized on $\widehat{\mathcal{V}}_n$. For example, if the entire function $\hat{h}_{j,n}(\cdot)$ is "far above" the infimum η^* for some j, then the corresponding inequality is removed from our subsequent inference. Similarly, for the remaining inequalities, we also remove subsets of $\mathcal{X}$ on which $\hat{h}_{j,n}(\cdot)$ is "far above" η^*. Such removal is helpful for attaining a more powerful test because the CSPA inference concentrates on the infimum η^* (recall (2.6)). In this selection step, the critical value $\hat{K}_n$ is defined as a "high" quantile (i.e., $\tilde{\gamma}_n \to 1$ slowly) of $\max_{1\leq j\leq J}\sup_{x\in\mathcal{X}}\hat{t}^*_{j,n}(x)$, so that the selection is conservative enough without leading to size distortion. Step 3 of the algorithm produces the upper confidence bound $\hat{\eta}_{n,1-\alpha}$, which is defined as the minimum of the upper envelope function $\hat{h}_{j,n}(\cdot) + n^{-1/2}\hat{k}_{n,1-\alpha}\hat{\sigma}_{j,n}(\cdot)$. Here, the critical value $\hat{k}_{n,1-\alpha}$ is computed based on the selected set $\widehat{\mathcal{V}}_n$. When $\widehat{\mathcal{V}}_n$ is smaller, $\hat{k}_{n,1-\alpha}$ is lower, which in turn leads to a lower value of the upper confidence bound $\hat{\eta}_{n,1-\alpha}$ and hence higher propensity of rejection. This explains more precisely how the adaptive selection may help improve the power of the test, particularly relative to the confidence bound computed under the least-favorable null (i.e., $\mathbb{E}[Y_{j,t}|X_t] = 0$) that corresponds to taking $\widehat{\mathcal{V}}_n$ as the whole set.

We are now ready to present the asymptotic theory that justifies the validity of the CSPA test described in Algorithm 1 above. As mentioned in the Introduction, our theory differs from Chernozhukov, Lee, and Rosen (2013) because we allow for general time series dependence and use a less restrictive notion of strong approximation. For ease of discussion, we collect the key ingredients of the theorem in the following two high-level assumptions. Below, we denote $\zeta_n \equiv \max_{1\leq j\leq m_n}\sup_{x\in\mathcal{X}}|p_j(x)|$ and $\zeta_n^L \equiv \sup_{x_1,x_2\in\mathcal{X},x_1\neq x_2}\|P(x_1)-P(x_2)\|/\|x_1-x_2\|$.

Assumption 2. *Suppose: (i) for each $j = 1,\ldots,J$, $h_j(\cdot)$ is a continuous function on a compact subset $\mathcal{X} \subseteq \mathbb{R}^{d_x}$; (ii) there exist sequences $(b^*_{j,n})_{n\geq 1}$ of m_n-dimensional constant vectors, with $m_n \to \infty$ at polynomial rate, such that*

$$\max_{1\leq j\leq J}\sup_{x\in\mathcal{X}} n^{1/2}\left|h_j(x) - P(x)^\top b^*_{j,n}\right| = o_p\left((\log n)^{-1}\right),$$

(iii) the eigenvalues of $Q_n \equiv n^{-1}\sum_{t=1}^n \mathbb{E}[P(X_t)P(X_t)^\top]$ and A_n are bounded from above and away from zero uniformly over n; (iv) $\|\widehat{Q}_n - Q_n\|_S = O_p(\delta_{Q,n})$ for $\delta_{Q,n} = o(m_n^{-1/2}(\log n)^{-1})$; (v) $\zeta_n m_n n^{-1/2} = o(1)$; and (vi) $\log(\zeta_n^L) = O(\log n)$.

The conditions in Assumption 2 are fairly standard for series estimation; see, for example, Andrews (1991b), Newey (1997), Chen (2007), Chernozhukov, Lee, and Rosen (2013), and Belloni, Chernozhukov, Chetverikov, and Kato (2015). In particular, condition (ii) specifies the precision for approximating the unknown function $h_j(\cdot)$ via approximating functions. This condition implicitly requires that the function $h_j(\cdot)$ is sufficiently smooth, for which well-known results are available

from numerical approximation theory. Condition (iv) imposes a mild convergence rate condition on $\widehat{Q}_n$, which can be verified under primitive conditions.[9]

Assumption 3. *For any sequence of integers $L_n = O(((\log n)^2 m_n^{1/2} \zeta_n^L)^{d_x})$ and any collection of uniformly bounded vectors $(\alpha_l)_{1 \leq l \leq L_n} \subseteq \mathbb{R}^{Jm_n}$, there exists a sequence of random variables $\widetilde{U}_n$ such that*

$$\left| \max_{1 \leq l \leq L_n} \alpha_l^\top \left(n^{-1/2} \sum_{t=1}^n u_t \otimes P(X_t) \right) - \widetilde{U}_n \right| = o_p\left((\log n)^{-1} \right), \tag{2.9}$$

and $\widetilde{U}_n$ has the same distribution as $\max_{1 \leq l \leq L_n} \alpha_l^\top \widetilde{N}_n$ for some generic Gaussian vector $\widetilde{N}_n \sim \mathcal{N}(0, A_n)$.

Assumption 3 is the key to the uniform functional inference underlying the CSPA test, and is nontrivial to establish. Note that our series estimation is based on the Jm_n-dimensional moment condition $\mathbb{E}[u_t \otimes P(X_t)] = 0$. The assumption above states that the maximum projection of the normalized growing-dimensional sample moment $n^{-1/2} \sum_{t=1}^n u_t \otimes P(X_t)$ can be approximated by $\widetilde{U}_n$, which has the same distribution as $\max_{1 \leq l \leq L_n} \alpha_l^\top \widetilde{N}_n$. In contrast, Yurinskii's coupling provides a strong approximation for the entire vector in Euclidean norm, namely,

$$\left\| n^{-1/2} \sum_{t=1}^n u_t \otimes P(X_t) - \widetilde{N}_n \right\| = o_p\left(\log(n)^{-1} \right), \tag{2.10}$$

which is clearly stronger than (2.9), but it invariably also demands more restrictive regularity conditions. An important part of our theoretical analysis (see Section 2.4) is to construct the coupling in (2.9) for general heterogeneous mixingales under substantially weaker conditions on the growth rate of m_n than those employed in Li and Liao (2020) for establishing Yurinskii's coupling in a similar time series setting.

A "cost" of using the weaker coupling condition (2.9), instead of (2.10), is that we do not have a strong Gaussian approximation for the entire t-statistic process

$$\left(\frac{n^{1/2} \hat{h}_{j,n}(x)}{\hat{\sigma}_{j,n}(x)} \right)_{1 \leq j \leq J, x \in \mathcal{X}},$$

which is required in Chernozhukov, Lee, and Rosen's (2013) intersection-bound theory as a high-level condition (see their Condition C.2). Consequently, we cannot directly invoke the theory from that prior work. Nevertheless, we show that under Assumption 3, one can still construct strong approximations for the supremum of the t-statistic process over all subsets of $\mathcal{X}$ (see Theorem

[9] See, for example, Lemma 2.2 of Chen and Christensen (2015) and Lemma B5 in the supplemental appendix of Li and Liao (2020).

A1 in the appendix), which turns out to be enough for establishing the validity of the testing procedure.

The asymptotic properties of the CSPA test are described by the following proposition.

Proposition 1. *Suppose that Assumptions 1, 2, and 3 hold. Then, the CSPA test at significance level $\alpha \in (0, 1/2)$ satisfies the following:*

(a) Under the null hypothesis with $\eta^* \geq 0$, the test has asymptotic size α, that is,

$$\limsup_{n\to\infty} \mathbb{P}\left(\hat{\eta}_{n,1-\alpha} < 0\right) \leq \alpha;$$

(b) Under the alternative hypothesis with $\eta^* < 0$, the test has asymptotic power one, that is,

$$\mathbb{P}\left(\hat{\eta}_{n,1-\alpha} < 0\right) \to 1.$$

Comments. (i) Part (a) of Proposition 1 shows that the CSPA test controls size under the null hypothesis. As is common for testing inequalities, the test may be conservative, that is, the asymptotic rejection probability may be less than α.

(ii) Part (b) shows that the test is consistent against fixed alternatives. As shown in our proof, this result follows from $\hat{\eta}_{n,1-\alpha} - \eta^* = O_p\left(\zeta_n m_n n^{-1/2}\right) = o_p(1)$. The proof can be straightforwardly adapted to show that the test is consistent against local alternatives with $\eta^* < 0$ drifting to zero at rate strictly slower than $\zeta_n m_n n^{-1/2}$. □

2.3 Confidence sets for the most superior forecasting method

The CSPA test described in the previous subsection concerns the comparison of a benchmark method with the other competing alternatives. In many applications, however, it may be a priori unclear which forecasting method should be chosen as the benchmark, and the empirical researcher may naturally experiment with different choices. This practice can be formalized as constructing a model confidence set for the most superior forecast, as we discuss in this subsection.

Formally, we define a partial order $\succeq$ between two forecasting methods indexed by j and k as

$$j \succeq k \iff \mathbb{E}\left[L(F_t^\dagger, F_{j,t})|X_t = x\right] \leq \mathbb{E}\left[L(F_t^\dagger, F_{k,t})|X_t = x\right] \text{ for all } x \in \mathcal{X}.$$

That is, the expected forecast loss of method j is less than that of method k across all conditioning states. The set of the most superior methods is then defined as

$$\mathcal{M} \equiv \{0 \leq j \leq J : j \succeq k \text{ for all } 0 \leq k \leq J\}. \tag{2.11}$$

Note that the CSPA null hypothesis with method j being the benchmark can then be written as $j \in \mathcal{M}$. Since the $\succeq$ order—which is defined using conditional expectation functions—is generally

not complete, the set $\mathcal{M}$ may be empty (i.e., there exists no method that weakly dominates all others).

It is interesting to contrast $\mathcal{M}$ with its unconditional special case, that is,

$$\mathcal{M}_U \equiv \{0 \leq j \leq J : \mathbb{E}[L(F_t^\dagger, F_{j,t})] \leq \mathbb{E}[L(F_t^\dagger, F_{k,t})] \text{ for all } 0 \leq k \leq J\}.$$

Since $\mathcal{M}_U$ relies on ordering the scalar-valued expected losses, it is guaranteed to be nonempty. It is also obvious that $\mathcal{M} \subseteq \mathcal{M}_U$ and, in general, we expect the inclusion to be strict. By imposing a stronger notion of dominance, $\mathcal{M}$ provides a refinement relative to its unconditional counterpart $\mathcal{M}_U$.

An Anderson–Rubin type confidence set for the most superior method can be constructed by inverting the CSPA test. We set

$$\begin{aligned}\widehat{\mathcal{M}}_{n,1-\alpha} = \{0 \leq j \leq J : &\text{ the } \alpha\text{-level CSPA test} \\ &\text{with method } j \text{ as the benchmark does not reject}\}.\end{aligned} \tag{2.12}$$

By the duality between tests and confidence sets, Proposition 1 directly implies that for each $j^* \in \mathcal{M}$,

$$\liminf_{n\to\infty} \mathbb{P}\left(j^* \in \widehat{\mathcal{M}}_{n,1-\alpha}\right) \geq 1 - \alpha.$$

We refer to $\widehat{\mathcal{M}}_{n,1-\alpha}$ as the *confidence set for the most superior (CSMS).*

We stress that the CSMS is distinct from the model confidence set (MCS) proposed by Hansen, Lunde, and Nason (2011) in two ways. First, the CSMS is based on conditional tests, while the MCS is based on unconditional ones; note that the unconditional test is a special case of the conditional test with X_t being empty. Second, the CSMS, $\widehat{\mathcal{M}}_{n,1-\alpha}$, is designed to cover each element j^* in $\mathcal{M}$, instead of the whole set $\mathcal{M}$. In contrast, the MCS provides coverage for the $\mathcal{M}_U$ set. Of course, this distinction is only relevant when $\mathcal{M}$ contains more than one method. While the unconditional expected losses of two distinct forecasting methods might be identical (as real numbers) and result in a non-singleton $\mathcal{M}_U$, it is hard to conceive a scenario in which two different forecast sequences share exactly the same conditional expected loss across all states in $\mathcal{X}$ (as functions).[10] For this reason, we argue that covering each most superior method serves essentially the same empirical goal as covering the whole set $\mathcal{M}$ in the context of conditional testing. The CSMS may thus be considered as the conditional extension of Hansen, Lunde, and Nason's (2011) MCS.

[10] See Zhu and Timmermann (2020) for additional discussions.

2.4 Strong approximation for the maximum projection

In this subsection, we establish a strong approximation that can be used to verify the high-level Assumption 3. Since this type of coupling result is of independent theoretical interest, and is broadly useful for other types of nonparametric uniform inference in time-series analysis, we present the theory in a general setting. This subsection may be skipped by readers who are mainly interested in the application of the CSPA test.

We now turn to the setting. Let $\|\cdot\|_q$ denote the L_q-norm of a random variable for $q \geq 1$. We consider an M_n-dimensional L_q-mixingale array $(X_{n,t})$ with respect to a filtration $(\mathcal{F}_{n,t})$. That is, $X_{n,t}$ satisfies, for each $1 \leq l \leq M_n$ and $k \geq 0$,

$$\|\mathbb{E}[X_{l,n,t}|\mathcal{F}_{n,t-k}]\|_q \leq \bar{c}_n \psi_k, \quad \|X_{l,n,t} - \mathbb{E}[X_{l,n,t}|\mathcal{F}_{n,t+k}]\|_q \leq \bar{c}_n \psi_{k+1}, \tag{2.13}$$

where $X_{l,n,t}$ denotes the lth component of $X_{n,t}$, and the constants $\bar{c}_n$ and ψ_k control the magnitude and the dependence of the $X_{n,t}$ variables, respectively. Recall that mixingales form a very general class of time series models, including martingale differences, linear processes, and various types of mixing and near-epoch dependent processes as special cases, and naturally allow for data heterogeneity; we refer the reader to Davidson (1994) for a comprehensive review. To verify Assumption 3, we can set $X_{n,t} = u_t \otimes P(X_t)$ and $M_n = Jm_n$.

Denote

$$S_n \equiv n^{-1/2} \sum_{t=1}^{n} X_{n,t}, \quad \Sigma_n \equiv Var(S_n).$$

For bounded M_n-dimensional vectors $(\alpha_l)_{1 \leq l \leq L_n}$, we aim to construct a sequence of random variables $\widetilde{U}_n$ such that $\widetilde{U}_n$ has the same distribution as $\max_{1 \leq l \leq L_n} \alpha_l^\top \widetilde{S}_n$ for $\widetilde{S}_n \sim \mathcal{N}(0, \Sigma_n)$ and

$$\left| \max_{1 \leq l \leq L_n} \alpha_l^\top S_n - \widetilde{U}_n \right| = o_p\left((\log n)^{-1}\right). \tag{2.14}$$

In a recent paper, Li and Liao (2020) establish a Yurinskii-type coupling in a similar mixingale setting, which has the form $\|S_n - \widetilde{S}_n\| = O_p(\delta_n)$ for some $\delta_n = o(1)$. The Yurinskii-type coupling is stronger than the coupling concept formulated in (2.14), but with a slower rate of convergence than the latter as discussed below. For ease of comparison, we impose the same regularity conditions as in Li and Liao (2020).

Assumption 4. *(i) For some $q \geq 3$, there exists a martingale difference array $X^*_{n,t}$ such that $\max_{1 \leq l \leq M_n} \|X^*_{l,n,t}\|_q \leq \bar{c}_n \psi^*$ for some finite constant $\psi^* > 0$ and*

$$\|S_n - S^*_n\| = O_p(\bar{c}_n M_n^{1/2} n^{-1/2})$$

for $S_n^* \equiv n^{-1/2}\sum_{t=1}^n X_{n,t}^*$*; (ii) the eigenvalues of* $\mathbb{E}\left[X_{n,t}^* X_{n,t}^{*\top}\right]$ *are uniformly bounded from above and away from zero; (iii) uniformly for any integer sequence* k_n *that satisfies* $n - o(1) \leq k_n \leq n$*,*

$$\left\| \sum_{t=1}^{k_n} \left(V_{n,t}^* - \mathbb{E}\left[V_{n,t}^*\right] \right) \right\|_S = O_p(r_n) \tag{2.15}$$

where $V_{n,t}^* \equiv n^{-1}\mathbb{E}\left[X_{n,t}^* X_{n,t}^{*\top} | \mathcal{F}_{n,t-1}\right]$ *and* $r_n = o(1)$ *is a real positive sequence; (iv) the largest eigenvalue of* Σ_n *is uniformly bounded.*

A few remarks on Assumption 4 are in order. Condition (i) directly imposes a martingale approximation for the mixingale array, which is a well-known and very useful property of mixingales.[11] This condition effectively reduces the task of constructing a strong approximation for S_n to a simpler one for the approximating martingale S_n^*. The other conditions are needed for analyzing the latter, which can also be established under primitive conditions. In particular, we note that condition (iii) can be generally derived by using a matrix law of large numbers, and it holds trivially with $r_n = 0$ if $V_{n,t}^*$ is deterministic (while allowed to be time-varying). As a concrete illustration of this assumption, we consider the following example.

Example (Martingale Approximation). Suppose that $X_{n,t}$ is formed as a linear process with the form $X_{n,t} = \sum_{|j|<\infty} \theta_j \varepsilon_{n,t-j}$ where $\varepsilon_{n,t}$ is a triangular array of independent variables with finite qth moments. Under the condition $\sum_{|j|<\infty} |j\theta_j| < \infty$, Assumption 4(i) can be verified with the approximating martingale difference array defined explicitly as $X_{n,t}^* = (\sum_{|j|<\infty} \theta_j)\varepsilon_{n,t}$, where the filtration is given by $\mathcal{F}_{n,t} = \sigma(\varepsilon_{n,s} : s \leq t)$. In this case, $V_{n,t}^* = n^{-1}(\sum_{|j|<\infty} \theta_j)\mathbb{E}\left[\varepsilon_{n,t}\varepsilon_{n,t}^\top\right](\sum_{|j|<\infty} \theta_j)^\top$ is deterministic and may be time-varying if $\mathbb{E}\left[\varepsilon_{n,t}\varepsilon_{n,t}^\top\right]$ depends on t. Condition (ii) is satisfied if $\mathbb{E}\left[\varepsilon_{n,t}\varepsilon_{n,t}^\top\right]$ and its inverse have bounded eigenvalues. Condition (iii) is automatically satisfied with $r_n = 0$. Although the volatility of the approximating martingale is deterministic, it is interesting to note that $X_{n,t}$ can have stochastic conditional volatility because $\mathbb{E}[X_{n,t}^2|\mathcal{F}_{n,t-1}]$ depends on the realizations of lagged innovations $(\varepsilon_s)_{s<t}$. □

We are now ready to state our main theorem on strong approximation.

Theorem 1. *Let* $(\alpha_l)_{1\leq l\leq L_n}$ *be* M_n*-dimensional real vectors with uniformly bounded Euclidean norm, and* $\widetilde{S}_n$ *be a generic* M_n*-dimensional random vector with distribution* $\mathcal{N}(0, \Sigma_n)$*. We set*

$$\begin{aligned} B_{1,n} &\equiv n^{-3/2}\sum_{t=1}^n \mathbb{E}\left[(\log L_n)^{3/2} \max_{1\leq l\leq L_n} \left(\mathbb{E}\left[(\alpha_l^\top X_{n,t}^*)^2 | \mathcal{F}_{n,t-1} \right] \right)^{3/2} + \max_{1\leq l\leq L_n} |\alpha_l^\top X_{n,t}^*|^3 \right], \\ B_{2,n} &\equiv \min\left\{ L_n^{1/q} \max_{1\leq t\leq n, 1\leq l\leq L_n} \|\alpha_l^\top X_{n,t}^*\|_q, M_n^{1/2} \right\} + (\log L_n)^{1/2}. \end{aligned}$$

[11] A sufficient condition is $\sum_{q\geq 0}\psi_q < \infty$, under which the martingale difference is defined as $X_{n,t}^* \equiv \sum_{s=-\infty}^{\infty}\{\mathbb{E}[X_{n,t+s}|\mathcal{F}_{n,t}] - \mathbb{E}[X_{n,t+s}|\mathcal{F}_{n,t-1}]\}$; see Lemma A4 in the supplemental appendix of Li and Liao (2020).

Suppose that $\bar{c}_n(\log L_n)^{1/2}M_n^{1/2}n^{-1/2}+B_{1,n}^{1/3}(\log L_n)^{2/3}+B_{2,n}r_n^{1/2}=o((\log n)^{-1})$ *and Assumption 4 holds. Then, there exists a sequence of random variables* $\widetilde{U}_n$ *such that* $\widetilde{U}_n$ *has the same distribution as* $\max_{1\leq l\leq L_n}\alpha_l^\top\widetilde{S}_n$ *and*

$$\left|\max_{1\leq l\leq L_n}\alpha_l^\top S_n-\widetilde{U}_n\right|=o_p\left((\log n)^{-1}\right).$$

Theorem 1 establishes the strong approximation for the maximum statistic $\max_{1\leq l\leq L_n}\alpha_l^\top S_n$. There are two components in the approximation error. The first is related to the martingale approximation and is of order $O_p(\bar{c}_n(\log L_n)^{1/2}M_n^{1/2}n^{-1/2})$. The other error term captures the strong approximation error for the maximum statistic $\max_{1\leq l\leq L_n}\alpha_l^\top S_n^*$ formed using the approximating martingale S_n^*, and has order $O_p(B_{1,n}^{1/3}(\log L_n)^{2/3}+B_{2,n}r_n^{1/2})$. In a setting with independent data, Chernozhukov, Chetverikov, and Kato (2014) prove a similar coupling result. Their proof heavily relies on symmetrization-based empirical process techniques that are not available in a general time series setting. We establish the coupling using very different techniques, which is a key theoretical contribution of the current paper.[12]

Theorem 1 can be used to verify Assumption 3. Since the theorem is somewhat complicated in its general form, we provide a corollary in a special case that is much easier to understand.

Corollary 1. *Let* $(\alpha_l)_{1\leq l\leq L_n}$ *and* $\widetilde{S}_n$ *be defined as in Theorem 1. Suppose that (i) Assumption 4 holds with* $\bar{c}_n=O(1)$ *and* $V_{n,t}^*$ *being deterministic; (ii)* $M_n\asymp n^a$ *for some* $a\in(0,1/3)$*; and (iii)* L_n *has polynomial growth as* $n\to\infty$*. Then, there exists a sequence of random variables* $\widetilde{U}_n$ *such that* $\widetilde{U}_n$ *has the same distribution as* $\max_{1\leq l\leq L_n}\alpha_l^\top\widetilde{S}_n$ *and*

$$\left|\max_{1\leq l\leq L_n}\alpha_l^\top S_n-\widetilde{U}_n\right|=o_p\left((\log n)^{-1}\right).$$

Comment. To apply this result in the context of Section 2.2, we set $M_n=Jm_n$ and note that m_n and L_n have polynomial growth as $n\to\infty$. Under the simple setting stated by Corollary 1, Assumption 3 is verified provided that $m_n=O(n^a)$ for some $a\in(0,1/3)$. In contrast, Li and Liao's (2020) Yurinskii-type coupling has the form $\|S_n-\widetilde{S}_n\|=O_p(m_n^{5/6}n^{-1/6})$, with the approximation error shrinking to zero when $m_n=O(n^a)$ for $a\in(0,1/5)$, which is notably more restrictive. □

[12]Without resorting to symmetrization, we develop a martingale technique consisting of two main steps. We first approximate the sequence $\max_{1\leq l\leq L_n}\alpha_l^\top S_n^*$ with $\max_{1\leq l\leq L_n}\alpha_l^\top S_n^+$, where S_n^+ is constructed as a martingale with its predictable quadratic covariation being equal to Σ_n. We then show that the maximum projection $\max_{1\leq l\leq L_n}\alpha_l^\top S_n^+$ can be strongly approximated by $\widetilde{U}_n$ by properly bounding the difference between their distribution functions. Li and Liao (2020) apply a similar proof strategy for the growing dimensional vector S_n^*, instead of the maximum projection $\max_{1\leq l\leq L_n}\alpha_l^\top S_n^*$. Focusing on the latter requires a more complicated and more precise calculation for the error bound, but allows us to derive the coupling result with a much faster convergence rate.

3 Monte Carlo study

In this section, we examine the finite-sample performance of the CSPA test in Monte Carlo experiments. Section 3.1 presents the setting and Section 3.2 reports the results. Additional findings are discussed in Supplemental Appendix SD.

3.1 The data generating process

We consider a setting with J conditional moments for $J = 1$, 3, or 5. The data are simulated according to the following data generating process (DGP):

$$\begin{aligned} Y_{j,t} &= 1 - a\, e^{-(X_t - c)^2} + u_{j,t}, && 1 \leq j \leq J, \\ X_t &= 0.5 X_{t-1} + \epsilon_t, && \text{with } \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, 0.75), \\ u_{j,t} &= \rho_u u_{j,t-1} + v_{j,t}, && \text{with } v_{j,t} \overset{iid}{\sim} \mathcal{N}(0, \sigma_v^2). \end{aligned}$$

We consider $c \in \{0, 0.5, 1\}$ and $\rho_u \in \{0, 0.4, 0.8\}$. We also set $\sigma_v^2 = 3(1 - \rho_u^2)$ so that the variance of $u_{j,t}$ is kept constant at 3.

A few remarks on this DGP are in order. First, the X_t process is a centered Gaussian AR(1) process with its variance normalized to unity. Second, the conditional expectation function $h_j(x) = \mathbb{E}[Y_{j,t}|X_t = x]$ has the form

$$h_j(x) = 1 - a\, e^{-(x-c)^2},$$

and it attains its minimum $1 - a$ at $x = c$. The corresponding unconditional expectation is

$$\mathbb{E}[Y_{j,t}] = \mathbb{E}[h_j(X_t)] = 1 - \frac{a\, e^{-c^2/3}}{\sqrt{3}}. \tag{3.1}$$

The c parameter plays a useful role in our experiments: Since the distribution of X_t is concentrated around zero, we can explore the effect of data density on the estimation accuracy for different parts of the $h_j(\cdot)$ function by varying c. Third, as we increase ρ_u from 0 to 0.8, the error series $u_{j,t}$ become more persistent, rendering time-series inference more difficult.

We impose the null and alternative hypotheses for CSPA as follows. The null hypothesis described in (2.5) is satisfied when $a = 1$, that is, $h_j(c) = 0$ and $h_j(x) > 0$ when $x \neq c$. On the other hand, when $a > 1$, the conditional moment violates the CSPA null hypothesis because $h_j(x) < 0$ when x falls in the $(c - \sqrt{\log a}, c + \sqrt{\log a})$ interval. The deviation of a from 1 thus quantifies the "distance" between the null and alternative hypotheses.

In additional to the proposed CSPA test, we also implement Hansen's (2005) USPA test for comparison. Although these tests concern different hypotheses, it is interesting to concretely

demonstrate how their difference manifests in the present numerical setting. From equation (3.1), we see that

$$\mathbb{E}\left[Y_{j,t}\right] \geq 0 \quad \text{if and only if} \quad a \leq \sqrt{3}e^{c^2/3}.$$

Therefore, when $1 < a \leq \sqrt{3}e^{c^2/3}$, the CSPA null hypothesis will be violated, whereas the USPA null hypothesis still holds. This corresponds to a situation in which a competing forecast model strictly outperforms the benchmark in certain regions of the conditioning state space, but, at the same time, underperforms the benchmark on average. By design, the USPA test should not reject, and the CSPA test will reject, providing useful additional diagnostic information. We consider three sample sizes, $n = 100$, 500, or 1,000. The simulation consists of 10,000 Monte Carlo replications.

Finally, we describe the implementation details of the CSPA test. The significance level is fixed at 5%. We set the approximating basis functions as Legendre polynomials evaluated on the rank-transformation of the conditioning variable X_t. That is, $P(X_t) = LP(2\,Rank(X_t) - 1)$, where $LP(z) = (1, z, 0.5(z^2 - 1), ...)$ denotes the Legendre polynomials, and $Rank(X_t)$ denotes the relative rank (i.e., empirical quantile) of X_t.[13] We use $m_n = \max\{4, \lfloor n^{1/5} \rfloor\}$ series terms, so that the series regression fits at least a cubic polynomial, and becomes increasingly more flexible in larger samples. To obtain estimates of the long-run variance $\widehat{\Omega}_n$, we consider both the Newey–West and pre-whitened HAC estimators. For the former, we use the Bartlett kernel (Newey and West (1987)) and, following Andrews (1991b), set the bandwidth to $\lfloor 0.75n^{1/3} \rfloor$. For the pre-whitened estimator that is described in detail in Supplemental Appendix SC, we pre-whiten the data using an autoregressive filter adaptively tuned via the AIC up to four lags.[14]

3.2 Results

To set the stage, we discuss results from the univariate setting (i.e., $J = 1$). We first examine the size property of the CSPA test, for which we impose the null hypothesis by setting $a = 1$. Table 1 reports the rejection rates of the test under the null for various values of ρ_u and c. The left and right panels report results based on the Newey–West HAC estimator and the pre-whitened estimator, respectively.

We summarize the results as follows. First, looking at the "small-sample" case with $n = 100$ (Panel A), we see that the test based on the standard HAC estimator controls size well when the

[13] Recall that the Legendre polynomials are orthogonal on $[-1, 1]$ under the Lebesgue measure. Our choice of approximating basis functions helps mitigating the potential issue of multicollinearity in the series regression involving many series terms, and it generally provides numerically stable results. Other types of orthogonal basis can be used similarly for this purpose.

[14] As in Andrews and Monahan (1992), the validity of the pre-whitened HAC estimator does not require the time series to actually follow a vector autoregressive model.

Table 1: Rejection Rates under the Null Hypothesis

	Newey–West HAC			Pre-whitened HAC		
	$\rho_u = 0.0$	$\rho_u = 0.4$	$\rho_u = 0.8$	$\rho_u = 0.0$	$\rho_u = 0.4$	$\rho_u = 0.8$
Panel A: Small-sample Case ($n = 100$)						
$c = 0.0$	0.017	0.030	0.098	0.018	0.029	0.056
$c = 0.5$	0.019	0.030	0.100	0.020	0.030	0.060
$c = 1.0$	0.016	0.028	0.090	0.017	0.027	0.062
Panel B: Medium-sample Case ($n = 500$)						
$c = 0.0$	0.005	0.010	0.039	0.005	0.008	0.014
$c = 0.5$	0.009	0.015	0.048	0.008	0.012	0.017
$c = 1.0$	0.001	0.003	0.020	0.001	0.002	0.008
Panel C: Large-sample Case ($n = 1{,}000$)						
$c = 0.0$	0.003	0.004	0.023	0.003	0.008	0.009
$c = 0.5$	0.009	0.010	0.033	0.009	0.010	0.015
$c = 1.0$	0.000	0.001	0.008	0.001	0.001	0.004

Note: This table presents rejection rates of the CSPA test under the univariate null hypothesis (i.e., $J = 1$ and $a = 1$). The test is implemented using either the Newey–West or the pre-whitened HAC estimator. We consider different data generating processes by varying the sample size n, residual autocorrelation ρ_u, and the location parameter c.

error terms are moderately persistent (i.e., $\rho_u = 0$ or 0.4). However, the test can be nontrivially oversized when $\rho_u = 0.8$. Indeed, the rejection rates are close to 10%, notably higher than the 5% nominal level. In contrast, tests based on the pre-whitened estimator show satisfactory size control even in the presence of high persistence. We note that the overrejection resulted from the standard HAC estimator is mainly a small-sample issue, which dampens as we increase the sample size. To be cautious, we focus only on results based on the pre-whitened HAC estimator in all our discussion below.

Second, we note that the CSPA test is generally conservative, as its rejection rate is often below the 5% nominal level. This is expected from the asymptotic theory, as we can see from (2.7) that the probability of type-I error is asymptotically bounded by α. The intuition for the conservativeness is as follows. In our simulation design, the inequality $h_j(x) \geq 0$ is binding at $x = c$. If this information were known a priori, we could compute the critical value by concentrating on the singleton $x = c$. However, in finite samples, we need to conservatively uncover the "binding region" using a nonparametric estimator. To the extent that this preliminary estimation is coarse, the resulting critical value is conservative.

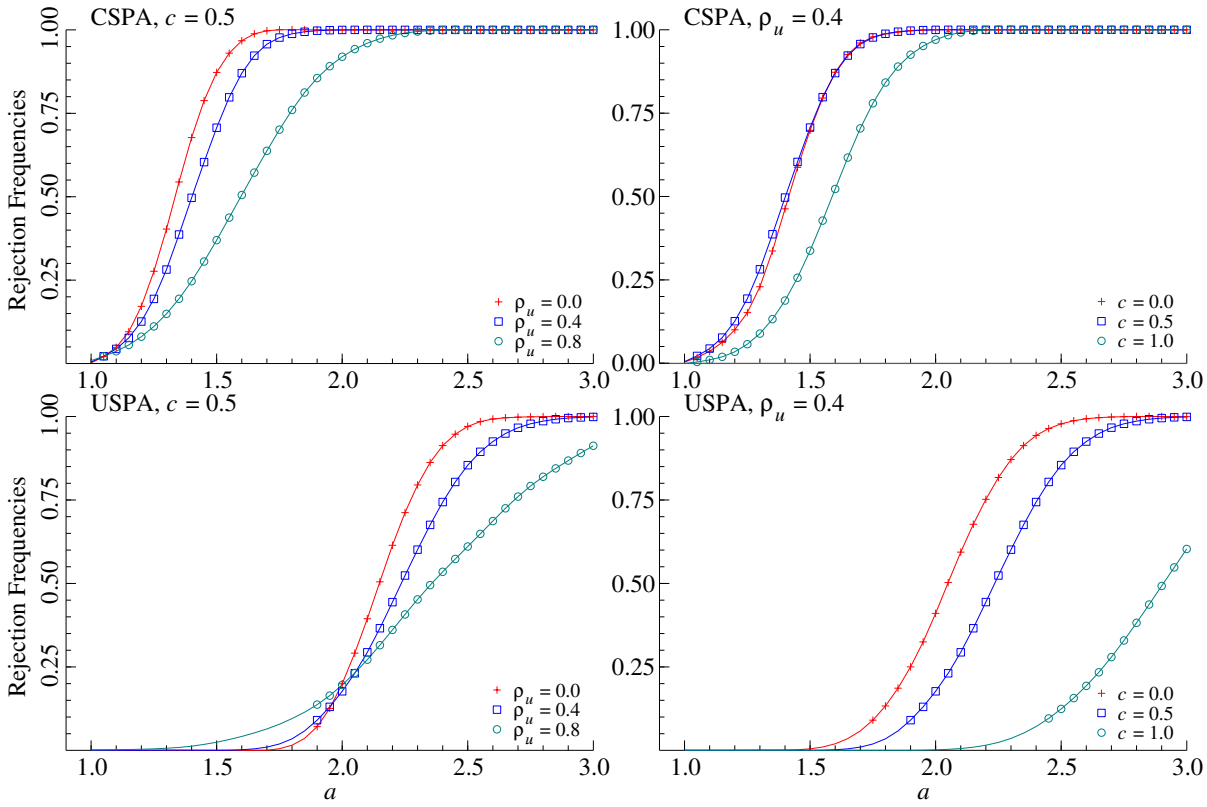
We next turn to power properties of the CSPA test, again for the univariate case with $J = 1$. As discussed above, the alternative hypothesis can be imposed by setting $a > 1$. In the top panels of Figure 1, we plot the CSPA test's rejection rates as functions of a while varying the ρ_u and c parameters in the DGP. For brevity, we only show results for the $n = 500$ case.

The top-left panel of Figure 1 plots the CSPA power curves for different levels of persistence quantified by the ρ_u parameter, while fixing $c = 0.5$. As a increases, the rejection rate approaches one, which is consistent with the asymptotic theory. In addition, we see that the test has higher power when the error terms are less persistent. On the top-right panel, we plot power curves for different c values while fixing $\rho_u = 0.4$. We see that the test rejects notably less frequently when $c = 1.0$, corresponding to the situation in which there are less data near the minimum of the conditional mean function, leading to less informative inference.

It is instructive to compare the power properties of CSPA test with those of Hansen's (2005) USPA test. The USPA power curves are plotted in the bottom panels of Figure 1, and computed under the same DGPs. We stress that this comparison should be interpreted cautiously, because CSPA and USPA tests are designed for different hypotheses. Specifically, the USPA null hypothesis is violated when $a > \sqrt{3}e^{c^2/3}$; the threshold is 1.73, 1.88, and 2.42 for the cases with $c = 0$, 0.5, and 1, respectively.

Looking at the USPA power curves, we see that the test essentially does not reject when a is less than the $\sqrt{3}e^{c^2/3}$ threshold; this is particularly evident in the bottom-right panel as we

Figure 1: Simulation Results: Power Curves



Note: This figure presents the power curves of the CSPA test (top row) and Hansen's (2005) USPA test (bottom row). We use sample size $n = 500$. In the left (resp. right) column we vary the ρ_u (resp. c) parameter while keeping the c (resp. ρ_u) parameter fixed. To highlight whether the value of a corresponds to the null or the alternative hypothesis, we signify the latter with a marker.

vary the value of c. When a exceeds this threshold, we observe increasingly more rejections, with the rejection rate approaching one as a becomes larger. Although the USPA test starts to have nontrivial power under its own alternative hypothesis, its rejection rate is notably lower than that of the CSPA test. This comparison thus concretely illustrates scenarios in which the conditional test can reveal useful information above and beyond its unconditional counterpart.

Finally, we present simulation results for the multivariate setting with $J = 3$ or 5. Similar to the univariate setting above, we impose the null hypothesis by setting $a = 1$, and set $a = 1.5$ to obtain the alternative. This alternative is chosen to have nondegenerate local power in view of Figure 1. Table 2 reports the rejection rates. From the table, we see that the CSPA test controls

Table 2: Simulation Results: Multivariate Test

	Null Hypothesis			Alternative Hypothesis		
	$\rho_u = 0$	$\rho_u = 0.4$	$\rho_u = 0.8$	$\rho_u = 0$	$\rho_u = 0.4$	$\rho_u = 0.8$
			Panel A: $J = 3$			
$c = 0.0$	0.003	0.006	0.016	0.992	0.910	0.507
$c = 0.5$	0.008	0.011	0.019	0.987	0.912	0.547
$c = 1.0$	0.001	0.002	0.010	0.705	0.500	0.271
			Panel B: $J = 5$			
$c = 0.0$	0.003	0.007	0.018	0.999	0.960	0.591
$c = 0.5$	0.009	0.011	0.023	0.998	0.965	0.648
$c = 1.0$	0.001	0.003	0.010	0.779	0.579	0.333

Note: This table presents rejection rates of the CSPA test when $J = 3$ or $J = 5$. We set $a = 1$ and 1.5 for the null and the alternative hypotheses, respectively. We consider different data generating processes of sample size $n = 500$, residual autocorrelation ρ_u, and the location parameter c.

size in the multivariate setting and is slightly more conservative. Meanwhile, the power of the test is higher as more conditional moment inequalities are violated under the alternative.

4 Empirical application on volatility forecast

As a first empirical application of the CSPA test, we consider the conditional evaluation of forecasting methods for realized volatility (RV). Ex-post measures of daily volatility based on high-frequency data, and the reduced-form modeling of their dynamics, are popularized by Andersen, Bollerslev, Diebold, and Labys (2003), whose seminal work has since spurred a large and burgeoning literature in financial econometrics. We apply the CSPA test on a collection of prominent forecasting methods. Section 4.1 introduces the forecasting models and reports baseline results from unconditional evaluation tests. Section 4.2 presents results from the CSPA test.

4.1 Volatility forecasting methods and their unconditional evaluation

Our analysis is based on the publicly available dataset from Bollerslev, Patton, and Quaedvlieg (2016), which contains daily realized volatility and other high-frequency measures for the S&P 500 index and 27 stocks in Dow Jones Industrial Average that are traded over the full sample period from April 1997 to December 2013. Realized volatility is computed as the sum of squared 5-minute returns within regular trading hours. We focus on one-day-ahead forecasts that are formed using rolling-window estimation with 1,000 daily observations. This results in over 3,000 daily forecasts for each series.[15]

We study six competing forecasting methods in total. The first is Corsi's (2009) HAR model, which is arguably the most popular model in the recent financial econometrics literature for volatility forecasting. This model is a restricted AR(22), in which common coefficients are imposed across "daily," "weekly," and "monthly" lags. We consider three alternative autoregressive specifications, including AR(1), AR(22), and an adaptive AR(22) model with LASSO-based variable selection.[16] Note that the HAR model is designed to capture the well-known long-memory feature of volatility. In this vein, we also include an ARFIMA$(1,d,0)$ model in our analysis, which is a classical alternative for capturing long memory via fractional integration. Lastly, we include one of many recent augmentations to HAR—the HARQ model proposed by Bollerslev, Patton, and Quaedvlieg (2016). HARQ allows the first autoregressive parameter to vary over time in response to the level of measurement error in the high-frequency estimate of daily RV. The level of measurement error is quantified by the daily realized quarticity (RQ), defined as the (scaled) sum of the 4th power of high-frequency returns. More precisely, the specifications of these forecast models are given below:

$$
\begin{aligned}
&\textbf{AR(1)} && RV_t = \phi_0 + \phi_1 RV_{t-1} + \epsilon_t, \\
&\textbf{AR(22)} && RV_t = \phi_0 + \textstyle\sum_{i=1}^{22} \phi_i RV_{t-i} + \epsilon_t, \\
&\textbf{HAR} && RV_t = \phi_0 + \phi_1 RV_{t-1} + \phi_2 RV_{t-1|t-5} + \phi_3 RV_{t-1|t-22} + \epsilon_t, \\
&\textbf{HARQ} && RV_t = \phi_0 + (\phi_1 + \phi_{1Q} RQ_{t-1}^{1/2}) RV_{t-1} + \phi_2 RV_{t-1|t-5} + \phi_3 RV_{t-1|t-22} + \epsilon_t, \\
&\textbf{ARFIMA} && (1-\mathbb{L})^d RV_t = \phi_0 + \phi_1 RV_{t-1} + \epsilon_t,
\end{aligned}
$$

where $RV_{t-1|t-k} = k^{-1} \sum_{i=1}^{k} RV_{t-i}$ and $\mathbb{L}$ denotes the lag operator.

To set the stage, we first conduct an unconditional comparison of these six methods using Hansen's (2005) USPA test. The test is implemented under two schemes: one-versus-one or one-

[15] The S&P500 realized volatility is based on futures data from Tick Data and ends in August 2013, resulting in a total of 4,096 observations. The individual stocks, whose data are obtained from the TAQ database, spans the full sample with 4,202 observations. See Bollerslev, Patton, and Quaedvlieg (2016) for details.

[16] The regularization parameter of the LASSO is obtained using a 10-fold cross-validation.

versus-all. Under one-versus-one, we compare each benchmark with one alternative for all model pairs. Under one-versus-all, we perform the USPA test to compare each benchmark model with all the other five competing models jointly. To mitigate the impact of rare but large volatility spikes, we focus on the relative forecasting error by using the following loss function

$$L(F_t^\dagger, F_{j,t}) = \frac{F_{j,t}}{F_t^\dagger} - \log\left(\frac{F_{j,t}}{F_t^\dagger}\right) - 1,$$

which is convex in the relative forecasting error $(F_{j,t} - F_t^\dagger)/F_t^\dagger$.

The tests are implemented separately for each asset. As a simple summary of these testing results, we report the number of assets (out of 28 in total) for which the USPA test (one-versus-one or one-versus-all) rejects at 5% significance level in Table 3.[17] From the top panel, we see considerable heterogeneity in the models' average loss: The average loss of the worst model, AR(1), is more than twice as high as that of the best model, HARQ. HAR appears to easily outperform AR(22)-LASSO. Interestingly, the latter adaptive method actually underperforms the AR(22) model with unrestricted coefficients.

We next turn to formal testing results. Under the one-versus-one scheme, we find that the USPA null hypothesis is almost never rejected when HARQ is the benchmark (see the fifth column). The only exception occurs when the competing method is ARFIMA, and the null hypothesis is rejected for 2 out of the 28 assets. But when ARFIMA is the benchmark, the associated USPA null is rejected for most of the assets against HARQ. Judged in a similar fashion, the other four methods can be ranked as follows: HAR, AR(22), AR(22)-LASSO, and AR(1). The one-versus-all tests, as shown in Panel C, more clearly confirms HARQ's superior performance. In particular, the joint test always rejects when each of the four underperforming models is used as the benchmark.

Overall, these unconditional evaluation results largely confirm prior findings in the literature. However, this does not preclude, for instance, the possibility that AR(1) can outperform HARQ and ARFIMA in *some* states of the world. The CSPA test is designed to investigate such issue, to which we now turn.

4.2 CSPA of volatility forecasts

We implement the CSPA test with the CBOE Volatility Index (VIX) as the conditioning state variable. The VIX is an option-based implied volatility measure, and is often deemed to be the "fear gauge" of investors. In our analysis below, we use each of the aforementioned models as a benchmark, and test whether $\mathbb{E}[Y_{j,t}|\text{VIX}_{t-1}] \geq 0$, where $Y_{j,t}$ denotes the loss differential between

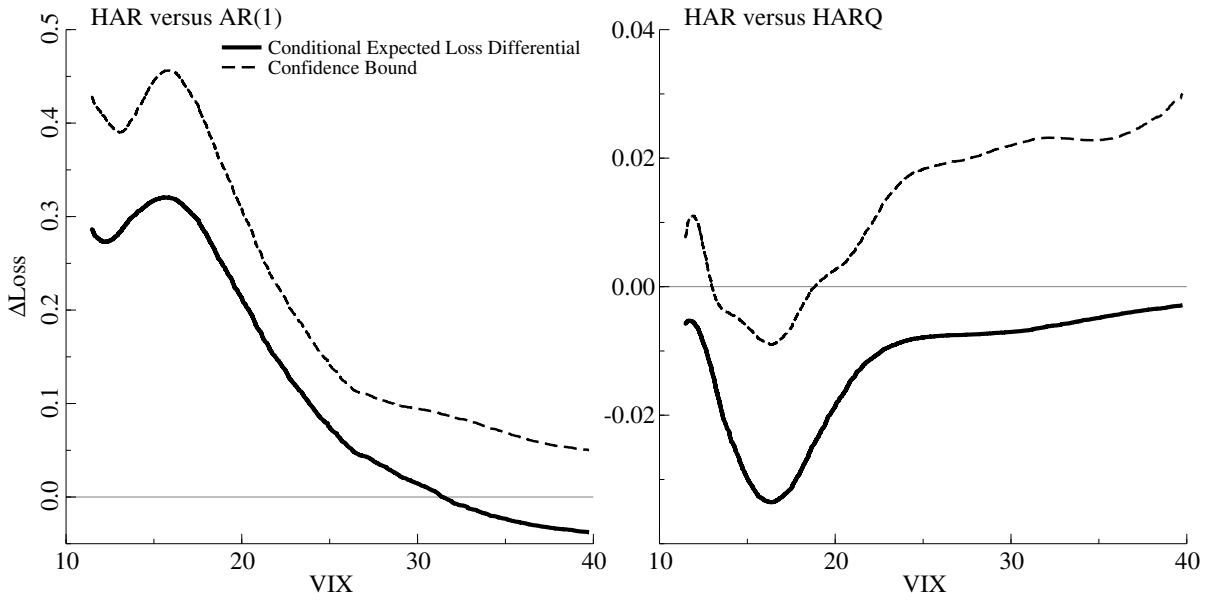
[17] We do not intend to interpret the testing results jointly across different assets.

Table 3: Unconditional Superior Predicative Ability for Volatility Forecasts

	Benchmark Methods					
	AR(1)	AR(22)	AR(22) LASSO	HAR	HARQ	ARFIMA
Panel A: Average Loss						
	0.401	0.229	0.298	0.213	0.185	0.198
Panel B: One-versus-one USPA tests against different competing methods						
AR(1)		0	3	0	0	0
AR(22)	28		27	0	0	0
AR(22) LASSO	25	0		0	0	0
HAR	28	26	28		0	1
HARQ	28	28	28	28		22
ARFIMA	28	27	28	26	2	
Panel C: One-versus-all USPA tests against all competing methods						
	28	28	28	28	2	22

Note: Panel A reports the average loss of each of the six forecasting methods, where the averaging is both over time and across assets. Panel B (resp. C) reports the number of assets for which the one-versus-one (resp. one-versus-all) USPA null hypothesis is rejected at 5% significance level.

Figure 2: Forecasting Volatility: One-versus-one CSPA Tests



Note: This figure plots the estimated conditional expected loss differential functions (solid), along with its 95% confidence bound (dashed), for the realized variance series of Johnson & Johnson, against the conditioning variable VIX. The HAR model is used as the benchmark, with the AR(1) (left panel) and the HARQ (right panel) as competing alternatives. A negative value of the conditional expected loss differential indicates that the HAR model is outperformed by the competing alternative. The CSPA test rejects the null hypothesis (i.e., HAR is weakly superior) if the confidence bound is below zero over some region of the conditioning state space.

the jth competing model and the benchmark. In particular, a non-rejection of the one-versus-all test suggests that the benchmark belongs to the CSMS. We follow the same implementation procedure using the pre-whitened HAC estimator as described in the simulation study.

To concretely illustrate how the CSPA test works in practice, we start with a case study for Johnson & Johnson (NYSE: JNJ). The evaluation scheme is one-versus-one: The benchmark is HAR and the competing alternative is either AR(1) or HARQ. In Figure 2, we plot the estimated conditional expected loss differential functions, along with the 95% upper confidence bounds given by $\hat{h}_{j,n}(\cdot)+n^{-1/2}\hat{k}_{n,1-\alpha}\hat{\sigma}_{j,n}(\cdot)$. Recall that the critical value $\hat{\eta}_{n,1-\alpha}$ of the CSPA test is the infimum of the latter function and the CSPA test rejects the null hypothesis if some part of this function is below zero.

The left panel illustrates the comparison between HAR and AR(1). As shown by the conditional

expected loss differential curve, the simple AR(1) forecast underperforms HAR when VIX is below 30, but appears to be more accurate in extremely volatile states. That being said, the CSPA null hypothesis is not rejected at the 5% level (because the confidence bound function is always above zero).

Turning to the right panel of Figure 2, we see that HARQ outperforms the benchmark HAR model not only on average but also uniformly across different states, in that the conditional expectation function is always below zero.[18] The former's better performance mainly stems from quiescent states (say, VIX $\approx$ 15), which is consistent with Bollerslev, Patton, and Quaedvlieg's (2016) finding that HARQ improves HAR by allowing for more persistence in quiet times, while sharply down-weighing past observations under crisis-like market conditions. Looking at the upper confidence bound, we see that HARQ's better performance is highly statistically significant when VIX ranges between 13.0 and 18.9, resulting in a rejection of the CSPA null hypothesis.

We now summarize all testing results. Panel A of Table 4 reports the number of assets, out of 28, for which the one-versus-one CSPA null hypothesis is rejected for each benchmark-competitor pair. In contrast to the unconditional testing results in Table 3, an interesting, and somewhat surprising, finding from the conditional tests pertains to the simple AR(1). Looking at the large rejection numbers in the first column of the table, we see clearly that AR(1) does not show any uniform superiority with respect to the other methods. But from the first row of the table, which shows the results with AR(1) being the competitor, we also see a nontrivial number of rejections, suggesting that the AR(1) model cannot be easily dominated by the others uniformly across different states, either. This finding mirrors the pattern seen in Figure 2: The AR(1) model, which has "short memory" and is fast-updating, can outperform those with long memory during extremely volatile periods. The moral is, when crisis hits, "amnesia" could be a bliss.

Panel B of Table 4 reports results from the one-versus-all CSPA tests. From here, we see more clearly that HARQ reigns supreme in the conditional evaluation, albeit not uniformly dominating. Specifically, HARQ belongs to the CSMS for 24 out of the 28 assets; in sharp contrast, ARFIMA is in the CSMS for only ten assets, and the other four forecasting methods are never in the CSMS. To provide further insight on the one-versus-all CSPA test, we make a visualization in Figure 3 for the case of Johnson & Johnson, where the left and right panels feature AR(1) and HARQ as benchmarks, respectively. In each panel, we plot the estimated conditional expected loss differential function for each competing alternative, their lower envelope, and the confidence bound of the latter. The left panel shows a clear CSPA rejection of the AR(1) benchmark, mainly

[18] A direct consequence is that the CSPA null hypothesis is not rejected if we instead take HARQ as the benchmark and HAR as the competing alternative.

Table 4: Conditional Superior Predicative Ability for Volatility Forecasts

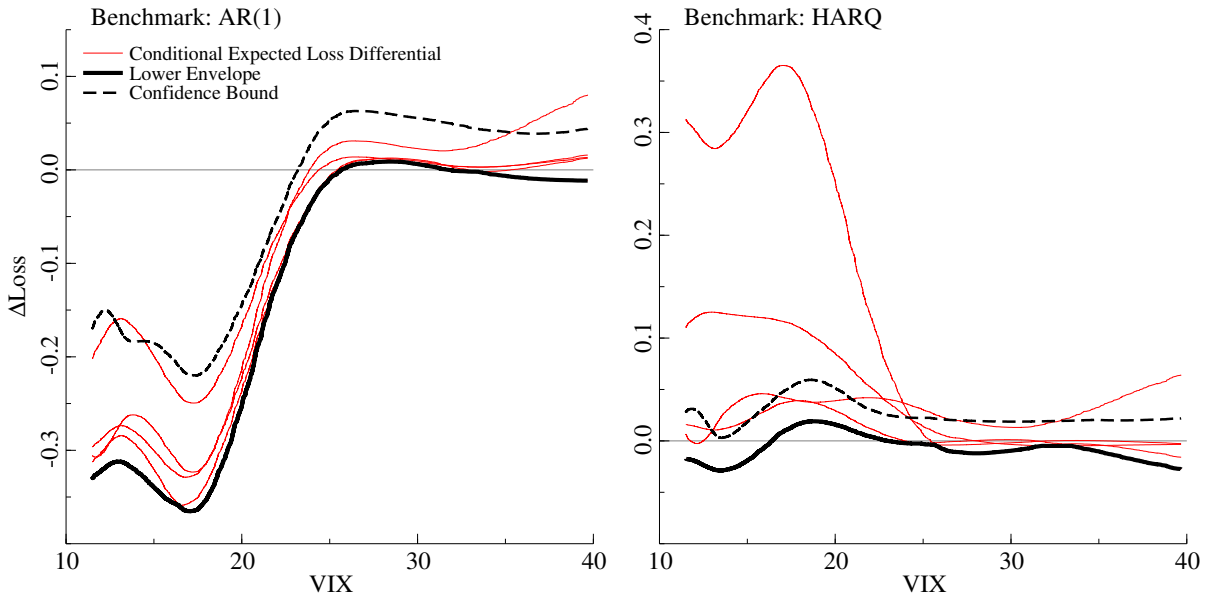
	Benchmark Methods					
	AR(1)	AR(22)	AR(22) LASSO	HAR	HARQ	ARFIMA
Panel A: One-versus-one CSPA tests against different competing models						
AR(1)		11	4	2	5	0
AR(22)	28		28	0	0	1
AR(22)+LASSO	28	18		0	0	0
HAR	28	24	28		0	2
HARQ	28	28	28	28		21
ARFIMA	28	28	28	28	2	
Panel B: One-versus-all CSPA tests against all competing models						
	28	28	28	28	4	18

Note: Panel A (resp. B) reports the number of assets, out of 28, for which the one-versus-one (resp. one-versus-all) CSPA null hypothesis is rejected at 5% significance level. Each column corresponds to a different benchmark.

due to the method's severe underperformance in low-VIX states. On the right panel, we see that, compared with HARQ, the best-performing competitor (which varies across different states) can reduce the conditional expected loss slightly in some states, but the reduction is statistically insignificant.

In summary, the application on volatility forecasting clearly highlights the usefulness of the CSPA test. We intentionally emphasize how nonparametric tools can provide richer diagnostic information regarding the state-dependent performance of different forecasting models, above and beyond conventional unconditional evaluation methods. As a by-product, we show that the recently developed HARQ model performs well not only on average but, quite often, also uniformly across different states for the loss function and the competing forecasting methods under consideration. That being said, the stringent CSPA criterion reveals that HARQ still needs improvement, and the

Figure 3: Forecasting Volatility: One-versus-all CSPA Tests



Note: In the left (resp. right) panel, we plot the estimated conditional expected loss differential functions, $\hat{h}_{j,n}(\cdot)$, using the AR(1) (resp. HARQ) model as the benchmark, and each of the other five models as the competing alternative. We also plot their lower envelope (solid dark), $\min_j \hat{h}_{j,n}(\cdot)$, along with its 95% upper confidence bound (dashed). The one-versus-all CSPA test rejects the null hypothesis if the confidence bound is below zero over some region of the conditioning state space.

search for uniformly superior volatility forecast should remain to be an important, and challenging, task on financial econometricians' research agenda.

5 Empirical application on inflation forecast

Our second application concerns inflation, which is notoriously difficult to forecast (Stock and Watson (2010); Faust and Wright (2013)) as evidenced by the fact that over extended periods of time a simple random walk forecast outperformed the official Greenbook inflation forecasts (Atkeson and Ohanian (2001), Faust and Wright (2009)). Meanwhile, in various academic and commercial applications, machine-learning methods have increasingly shown their potential for improving classical prediction methods. In a recent paper, Medeiros, Vasconcelos, Veiga, and Zilberman (2019) experiment with a variety of machine-learning methods including shrinkage method and random forests, among others. In this section, we apply the CSPA test to examine the conditional performance of

these forecasts, along with traditional ones. This is a particularly relevant application of the CSPA test in a macroeconometric context in light of the considerable instability in the performance of inflation forecasting methods (Stock and Watson (2009, 2010)).

We apply the CSPA test to evaluate eight inflation forecasting methods, which are selected from those constructed by Medeiros, Vasconcelos, Veiga, and Zilberman (2019) for the one-month-ahead and twelve-month-ahead forecasts of Consumer Price Index (CPI).[19] This includes four traditional methods: random walk, AR(1), Bayesian vector-autoregression (BVAR), and a factor model. The other four methods rely on machine-learning techniques, such as random-forest regression (RF-OLS), LASSO, elastic net (ElNet), and bagging. Specifically, RF-OLS implements a standard linear regression based on variables selected by random forest (Breiman (2001)); LASSO is proposed by Tibshirani (1996) and has been used in inflation forecasting by Bai and Ng (2008); ElNet combines the classical ridge regression with the LASSO method (Zou and Hastie (2005)); Bagging, or bootstrap aggregation (Breiman (1996)), is first used to forecast inflation by Inoue and Kilian (2008).[20] The evaluation sample period is between January 1990 and December 2015, consisting of 312 monthly observations in total. For brevity, we only focus on one-versus-all tests conducted jointly for all competing forecasts. All tests below are based on the quadratic loss function and are at the 5% significance level.

When implementing the CSPA test on these inflation forecasts, it is a priori unclear which conditioning state variable would be the most revealing, largely because of the intrinsic difficulty in inflation forecasting and many potentially relevant macroeconomic variables. We thus consider a number of important—and conceptually distinct—conditioning variables, including: average inflation over the past three months (INFL), industrial production growth (ΔIP), unemployment rate (UR), 12-month macro uncertainty (MU), economic policy uncertainty (EPU), equity market volatility (EMV), the VIX, and two measures of credit spread including the Baa less 10-year Treasury yield (CS) and Gilchrist and Zakrajšek's (2012) spread (GZ).[21]

[19] We are grateful to Marcelo C. Medeiros for kindly providing their forecasts and data. The forecasts are constructed using a rolling window of 360 months. The estimation sample starts from January 1960 and is based on the 2016 vintage of the Fred-MD database (McCracken and Ng (2016)). For brevity, we refer the reader to Appendix B of Medeiros, Vasconcelos, Veiga, and Zilberman (2019) for the complete description of their models.

[20] These four methods are selected because they have the smallest average mean-square-errors among all 18 methods studied by Medeiros, Vasconcelos, Veiga, and Zilberman (2019). In results not reported here, we instead analyze the full list. But the larger-scale study does not change our main empirical findings.

[21] The INFL, ΔIP, UR, and CS series are obtained from McCracken and Ng's (2016) FRED-MD database. The MU index is proposed by Jurado, Ludvigson, and Ng (2015) and retrieved from `www.sydneyludvigson.com`. The EPU and EMV indexes are constructed by Baker, Bloom, and Davis (2016), and retrieved from `www.policyuncertainty.com`. GZ was proposed by Gilchrist and Zakrajšek's (2012), and retrieved from `people.bu.edu/sgilchri`.

Table 5: USPA and CSPA Tests for Inflation Forecasts

Panel A: One-month-ahead Forecasts			CSPA								
	RMSE	USPA	CS	MU	VIX	INFL	ΔIP	UR	EMV	GZ	EPU
RF-OLS	0.65										
LASSO	0.69			✗							✗
ElNet	0.69		✗								
Bagging	0.69		✓								✗
BVAR	0.74	✓		✓			✗			✗	
Factor	0.76	✓	✓	✓	✓	✗	✗	✓	✓	✓	✗
AR	0.81	✓		✓	✓	✓	✓		✗	✗	
RW	1.00	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Panel B: Twelve-month-ahead Forecasts			CSPA								
	RMSE	USPA	CS	MU	VIX	INFL	ΔIP	UR	EMV	GZ	EPU
RF-OLS	0.50										
LASSO	0.53			✗							✗
ElNet	0.53			✗							
Bagging	0.54		✓					✗			✗
BVAR	0.55	✓		✓			✗			✗	
AR	0.56	✓		✓	✓	✓	✓		✗	✗	
Factor	0.61	✓	✓	✓	✓	✗	✗	✓	✓	✓	✗
RW	1.00	✓	✓	✓	✓	✓	✓	✓	✓	✓	

Note: The first column reports the relative mean-squared-error (RMSE) of each forecast method's average quadratic loss with respect to that of the random walk model. The second column reports rejections of the one-versus-all USPA hypothesis. The next nine columns present one-versus-all CSPA rejections, using the nine conditioning variables separately. For each of the entries, the ✓ symbol denotes a rejection at the 5% familywise-error (FWE) level, applying the Bonferroni–Hochberg p-value bounds row by row. The ✗ symbol denotes a rejection at the 5% nominal level, but not at the 5% FWE level. Forecast methods are sorted based on their RMSEs displayed in the first column.

Table 5 presents results of the USPA and CSPA tests for both forecasting horizons. The first two columns summarize the unconditional performance of the eight forecasting methods. The first column reports each method's relative mean-squared-error (RMSE), defined as the ratio between its average quadratic loss and that of the random-walk forecast. Note that the methods are ordered according to their RMSEs in the table. On the second column, we use the check mark to signify the one-versus-all USPA rejection for the method in each row (as the benchmark) against all of the other seven methods (as the competitors). From here, we see that the four machine-learning methods cannot be "separated" by the USPA test, in that the test does not reject those benchmarks. This is hardly surprising, because LASSO, ElNet, and Bagging have essentially the same average losses. That being said, the USPA test does provide strong statistical evidence for the underperformance of the "traditional" forecasting methods, such as AR(1).

We further turn to the CSPA test to examine the conditional performance of these forecasting methods, particularly the ones based on machine learning. In columns 3–11, we present the rejection decisions of the CSPA tests based on the aforementioned nine conditioning variables separately (i.e., nine one-versus-all CSPA tests are performed for each benchmark).[22] To guard against the issue regarding multiple testing, we also control the family-wise error (FWE) at 5% level by applying the Bonferroni–Hochberg method (Hochberg (1988)). We use the check and cross marks to signify rejections with or without FWE control, respectively; note that the former implies the latter.

Like the unconditional test, the conditional test also clearly shows the better performance of the machine-learning methods relative to the "traditional" ones, evidenced by the larger number of rejections when the latter methods are used as benchmarks. Indeed, with AR, Factor, or RW as the benchmark, the CSPA null hypothesis is almost always rejected for the various conditioning variables, and many of the rejections remain to hold under the (more stringent) FWE control. The BVAR benchmark is rejected less frequently, for which the key evidence stems from conditioning on macro uncertainty. More interestingly, in contrast to the unconditional test, the conditional test also helps distinguishing the four machine-learning methods. The most robust finding concerns the bagging method, as it is rejected by conditioning on the CS credit spread for both one-month-ahead and twelve-month-ahead forecasts, even if we control for the FWE across all conditioning variables. Meanwhile, the good performance of the RF-OLS method is notable, in that it not only has the smallest average loss, but also exhibits uniform conditional superiority as postulated by

[22] We do not use all conditioning variables jointly because of the well-known "curse of dimensionality" issue for nonparametric procedures (i.e., the rates of convergence of nonparametric estimators decline substantially when the dimension of X_t is high).

the CSPA null hypothesis for the broad variety of conditioning variables under consideration.

Overall, the analysis above further demonstrates the usefulness of the CSPA test in a macroeconometric context. Unlike our previous application on volatility, this example presents a concrete empirical scenario in which unconditional evaluation is silent on the relative performance of certain forecasting methods (though it does an adequate job of signifying the worst performing methods), and the CSPA test—by imposing a more stringent uniform requirement under the null—can be used to discriminate among unconditionally indistinguishable methods.

6 Conclusion

Motivated by the ever-increasing variety of forecasting methods in many areas of research, we introduce a new test for conditional superior predictive accuracy, or CSPA. This test examines the conditional state-dependent performance of competing forecasting methods, and imposes a more stringent uniform weak dominance requirement on the benchmark. Two empirical applications from financial and macroeconomic settings demonstrate the discriminating power of the CSPA test relative to its conventional unconditional counterpart. Econometrically, this is achieved by introducing and extending recently developed theoretical tools for the uniform nonparametric inference in the time-series setting, as CSPA concerns functional inequalities defined by conditional expected loss differentials. To the best of our knowledge, this is the first application of conditional-moment-inequality methods in time-series econometrics, and the theoretical tools developed here are broadly useful for other types of inference problems involving partial identification and dependent data.

Although the CSPA test provides an evaluation among the competing forecasts, it does not directly address the important question of how to improve the underlying forecasting methods. Possible strategies might be devised in a similar spirit as the switching rule considered by Giacomini and White (2006), or more generally, by properly "averaging" forecasting methods that are deemed superior by the test. These further investigations are beyond the scope of the present paper, and are left for future research.

Supplementary Data. Supplementary data are available at *Review of Economic Studies online.* The replication package is available at https://dx.doi.org/dx.doi.org/10.5281/zenodo.4884813.

Data Availability Statement. The data and code underlying this research is available on Zenodo at https://dx.doi.org/10.5281/zenodo.4884813.

References

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003): "Modeling and Forecasting Realized Volatility," *Econometrica*, 71(2), 529–626.

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

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

Andrews, D. W. K., and J. C. Monahan (1992): "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60(4), 953–966.

Atkeson, A., and L. E. Ohanian (2001): "Are Phillips Curves Useful for Forecasting Inflation?," *Federal Reserve bank of Minneapolis Quarterly Review*, 25(1), 2–11.

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

Baker, S. R., N. Bloom, and S. J. Davis (2016): "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, 131(4), 1593–1636.

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

Bollerslev, T., A. J. Patton, and R. Quaedvlieg (2016): "Exploiting the Errors: A Simple Approach for Improved Volatility Forecasting," *Journal of Econometrics*, 192(1), 1–18.

Bollerslev, T., A. J. Patton, and R. Quaedvlieg (2018): "Modeling and Forecasting (un) Reliable Realized Covariances for more Reliable Financial Decisions," *Journal of Econometrics*, 207(1), 71–91.

Breiman, L. (1996): "Bagging Predictors," *Machine learning*, 24(2), 123–140.

——— (2001): "Random Forests," *Machine learning*, 45(1), 5–32.

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

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

Chernozhukov, V., D. Chetverikov, and K. Kato (2014): "Gaussian Approximation of Suprema of Empirical Processes," *Annals of Statistics*, 42(4), 1564–1597.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

Corsi, F. (2009): "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 7(2), 174–196.

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

Diebold, F. X. (2015): "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests," *Journal of Business & Economic Statistics*, 33(1), 1–9.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3), 253–263.

Faust, J., and J. H. Wright (2009): "Comparing Greenbook and Reduced Form Forecasts using a Large Realtime Dataset," *Journal of Business & Economic Statistics*, 27(4), 468–479.

——— (2013): "Forecasting Inflation," in *Handbook of economic forecasting*, vol. 2, pp. 2–56. Elsevier.

Giacomini, R., and B. Rossi (2010): "Forecast Comparisons in Unstable Environments," *Journal of Applied Econometrics*, 25(4), 595–620.

——— (2016): "Model Comparisons in Unstable Environments," *International Economic Review*, 57(2), 369–392.

Giacomini, R., and H. White (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74(6), 1545–1578.

Gilchrist, S., and E. Zakrajšek (2012): "Credit spreads and business cycle fluctuations," *American Economic Review*, 102(4), 1692–1720.

Hansen, P. R. (2005): "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23(4), 365–380.

Hansen, P. R., A. Lunde, and J. M. Nason (2011): "The Model Confidence Set," *Econometrica*, 79(2), 453–497.

Herbst, E., and F. Schorfheide (2012): "Evaluating DSGE model forecasts of comovements," *Journal of Econometrics*, 171(2), 152 – 166.

Hochberg, Y. (1988): "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, 75(4), 800–802.

Inoue, A., and L. Kilian (2008): "How Useful is Bagging in Forecasting Economic Time Series? A Case Study of US Consumer Price Inflation," *Journal of the American Statistical Association*, 103(482), 511–522.

Jurado, K., S. C. Ludvigson, and S. Ng (2015): "Measuring uncertainty," *American Economic Review*, 105(3), 1177–1216.

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

McCracken, M. W., and S. Ng (2016): "FRED-MD: A Monthly Database for Macroeconomic Research," *Journal of Business & Economic Statistics*, 34(4), 574–589.

Medeiros, M. C., G. Vasconcelos, A. Veiga, and E. Zilberman (2019): "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods," *Journal of Business & Economic Statistics, forthcoming*.

Molinari, F. (2019): "Econometrics with Partial Identification," Discussion paper.

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

Newey, W. K., and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.

Rossi, B. (2013): "Chapter 21 - Advances in Forecasting under Instability," in *Handbook of Economic Forecasting*, ed. by G. Elliott, and A. Timmermann, vol. 2 of *Handbook of Economic Forecasting*, pp. 1203 – 1324. Elsevier, Amsterdam, Netherlands.

Rossi, B., and T. Sekhposyan (2010): "Have Economic Models' Forecasting Performance for US Output Growth and Inflation Changed Over Time, and When?," *International Journal of Forecasting*, 26(4), 808–835.

Stock, J. H., and M. W. Watson (2009): "Phillips Curve Inflation Forecasts," in *Understanding Inflation and the Implications for Monetary Policy*, ed. by J. Fuhrer, Y. Kodrzycki, J. Little, and G. Olivei. MIT Press, Cambridge.

——— (2010): "Modeling Inflation after the Crisis," Discussion paper, National Bureau of Economic Research.

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

West, K. D. (1996): "Asymptotic Inference about Predictive Ability," *Econometrica*, 64(5), 1067–1084.

White, H. (2000): "A Reality Check for Data Snooping," *Econometrica*, 68(5), 1097–1126.

Zhu, Y., and A. Timmermann (2020): "Can Two Forecasts Have the Same Conditional Expected Accuracy?," Discussion paper.

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

A Proofs

In this appendix, we prove the theoretical results in the main text. To highlight our new technical contributions, we demonstrate the main steps of our proofs here, and relegate additional technical details to the Supplemental Appendix. Throughout this appendix, we use K to denote a generic positive finite constant, which may vary from line to line.

A.1 Proof of Proposition 1

The proof of Proposition 1 shares some similar steps with Chernozhukov, Lee, and Rosen's (2013) proofs of their Theorem 2 and Lemma 5. To avoid repetition, we only highlight the key difference here and provide a detailed step-by-step proof in the Supplemental Appendix. An important difference stems from the fact that we rely on a strong approximation condition in the form of

Assumption 3, instead of the (stronger) Yurinskii-type coupling used in Lemma 5 of Chernozhukov, Lee, and Rosen (2013), which in turn is needed to verify their high-level Condition C.2 for the uniform strong approximation of the t-statistic process. A key step of our analysis is to establish Theorem A1 below. We denote

$$\sigma_{j,n}(x) \equiv \sqrt{P(x)^\top Q_n^{-1} A_n(j,j) Q_n^{-1} P(x)},$$

where $A_n(j,j)$ is the jth $m_n \times m_n$ diagonal block of the matrix A_n. For two random variables X and Y, we write $X \overset{d}{=} Y$ if they have the same distribution.

Theorem A1. *Let V_n be any subset of $\mathcal{V} = \{(j,x) : 1 \le j \le J \text{ and } x \in \mathcal{X}\}$ and $\widetilde{N}_n$ be a generic sequence of random variables such that $\widetilde{N}_n \sim \mathcal{N}(0, A_n)$. Suppose that Assumptions 2 and 3 hold. Then, there exist random sequences $\widetilde{U}_{1,n}$, $\widetilde{U}_{2,n}$, and $\widetilde{U}_{3,n}$ such that*

$$\widetilde{U}_{1,n} \overset{d}{=} \widetilde{U}_{2,n} \overset{d}{=} \sup_{(j,x)\in V_n} \frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)} \widetilde{N}_{j,n}, \quad \widetilde{U}_{3,n} \overset{d}{=} \sup_{(j,x)\in V_n} \left| \frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)} \widetilde{N}_{j,n} \right|,$$

and

$$\begin{aligned}
\sup_{(j,x)\in V_n} \frac{n^{1/2}(\hat{h}_{j,n}(x) - h_j(x))}{\sigma_{j,n}(x)} - \widetilde{U}_{1,n} &= o_p\left((\log n)^{-1}\right), \\
\sup_{(j,x)\in V_n} \frac{n^{1/2}(h_j(x) - \hat{h}_{j,n}(x))}{\sigma_{j,n}(x)} - \widetilde{U}_{2,n} &= o_p\left((\log n)^{-1}\right), \\
\sup_{(j,x)\in V_n} \left| \frac{n^{1/2}(h_j(x) - \hat{h}_{j,n}(x))}{\sigma_{j,n}(x)} \right| - \widetilde{U}_{3,n} &= o_p\left((\log n)^{-1}\right),
\end{aligned}$$

where $\widetilde{N}_{j,n}$ is the m_n-dimensional subvector defined implicitly by the partition $\widetilde{N}_n^\top = (\widetilde{N}_{1,n}^\top, \ldots, \widetilde{N}_{J,n}^\top)$.

Proof. Step 1. In this step, we show that

$$\sup_{(j,x)\in\mathcal{V}} \left| \frac{n^{1/2}(\hat{h}_{j,n}(x) - h_j(x))}{\sigma_{j,n}(x)} - \frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)} n^{-1/2} \sum_{t=1}^n P(X_t) u_{j,t} \right| = o_p\left((\log n)^{-1}\right). \tag{A.1}$$

Denote $h_{j,n}^*(\cdot) = P(\cdot)^\top b_{j,n}^*$, with $b_{j,n}^*$ given by Assumption 2. By Assumption 2, $\sigma_{j,n}(\cdot)$ is bounded away from zero. We then deduce that

$$\sup_{(j,x)\in\mathcal{V}} \left| \frac{n^{1/2}(h_j(x) - h_{j,n}^*(x))}{\sigma_{j,n}(x)} \right| \le K \sup_{(j,x)\in\mathcal{V}} \left| n^{1/2}(h_j(x) - h_{j,n}^*(x)) \right| = o_p\left((\log n)^{-1}\right). \tag{A.2}$$

Observe that

$$\hat{b}_{j,n} - b_{j,n}^* = \widehat{Q}_n^{-1} \left(n^{-1} \sum_{t=1}^n P(X_t) u_{j,t} + n^{-1} \sum_{t=1}^n P(X_t) (h_j(X_t) - h_{j,n}^*(X_t)) \right). \tag{A.3}$$

Therefore, by the triangle inequality and the Cauchy–Schwarz inequality, we have uniformly in x,

$$\begin{aligned}
&\left|\frac{n^{1/2}(\hat{h}_{j,n}(x)-h^*_{j,n}(x))}{\sigma_{j,n}(x)}-\frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)}n^{-1/2}\sum_{t=1}^n P(X_t)\,u_{j,t}\right| \\
&\quad\le\left|\frac{P(x)^\top(\widehat{Q}_n^{-1}-Q_n^{-1})}{\sigma_{j,n}(x)}n^{-1/2}\sum_{t=1}^n P(X_t)\,u_{j,t}\right| \\
&\qquad+\left|\frac{P(x)^\top\widehat{Q}_n^{-1}}{\sigma_{j,n}(x)}n^{-1/2}\sum_{t=1}^n P(X_t)\,(h_j(X_t)-h^*_{j,n}(X_t))\right| \\
&\quad\le K\left\|\widehat{Q}_n^{-1}-Q_n^{-1}\right\|_S\left\|n^{-1/2}\sum_{t=1}^n P(X_t)\,u_{j,t}\right\| \\
&\qquad+K\left\|\widehat{Q}_n^{-1}n^{-1/2}\sum_{t=1}^n P(X_t)\,(h_j(X_t)-h^*_{j,n}(X_t))\right\|. \qquad \text{(A.4)}
\end{aligned}$$

It is easy to see that $\|n^{-1/2}\sum_{t=1}^n P(X_t)\,u_{j,t}\|=O_p(m_n^{1/2})$. Then, by Assumption 2,

$$\left\|\widehat{Q}_n^{-1}-Q_n^{-1}\right\|_S\left\|n^{-1/2}\sum_{t=1}^n P(X_t)\,u_{j,t}\right\|=O_p\left(\delta_{Q,n}m_n^{1/2}\right)=o_p\left((\log n)^{-1}\right). \qquad \text{(A.5)}$$

In addition, we note that

$$\begin{aligned}
&\left\|\widehat{Q}_n^{-1}n^{-1/2}\sum_{t=1}^n P(X_t)\,(h_j(X_t)-h^*_{j,n}(X_t))\right\|^2 \\
&\quad\le O_p(1)\sum_{t=1}^n\left(h_j(X_t)-h^*_{j,n}(X_t)\right)^2=o_p\left((\log n)^{-2}\right). \qquad \text{(A.6)}
\end{aligned}$$

By (A.5) and (A.6), the majorant side of (A.4) can be further bounded by $o_p((\log n)^{-1})$. This estimate and (A.2) imply (A.1) as claimed.

Step 2. Recall that ζ_n^L is the Lipschitz coefficient for the $P(\cdot)$ function. By the triangle inequality and the Cauchy–Schwarz inequality, we have uniformly for $x_1,x_2\in\mathcal{X}$,

$$|\sigma_{j,n}(x_1)-\sigma_{j,n}(x_2)|=\frac{\left|\sigma^2_{j,n}(x_1)-\sigma^2_{j,n}(x_2)\right|}{\sigma_{j,n}(x_1)+\sigma_{j,n}(x_2)}\le K\zeta_n^L\|x_1-x_2\|. \qquad \text{(A.7)}$$

We then observe, for any $j \in \{1, \ldots, J\}$ and $x_1, x_2 \in \mathcal{X}$,

$$\begin{aligned}
&\left| \left(\frac{P(x_1)}{\sigma_{j,n}(x_1)} - \frac{P(x_2)}{\sigma_{j,n}(x_2)} \right)^\top Q_n^{-1} n^{-1/2} \sum_{t=1}^n P(X_t)\, u_{j,t} \right| \\
&\quad \leq \|P(x_1) - P(x_2)\| \, \frac{\left\| Q_n^{-1} n^{-1/2} \sum_{t=1}^n P(X_t)\, u_{j,t} \right\|}{\sigma_{j,n}(x_1)} \\
&\qquad + \frac{|\sigma_{j,n}(x_1) - \sigma_{j,n}(x_2)|}{\sigma_{j,n}(x_1)} \frac{\|P(x_2)\|}{\sigma_{j,n}(x_2)} \left\| Q_n^{-1} n^{-1/2} \sum_{t=1}^n P(X_t)\, u_{j,t} \right\| \\
&\quad \leq K \zeta_n^L \|x_1 - x_2\| \left\| Q_n^{-1} n^{-1/2} \sum_{t=1}^n P(X_t)\, u_{j,t} \right\|, \qquad \text{(A.8)}
\end{aligned}$$

where the first inequality is by the triangle inequality and the Cauchy–Schwarz inequality, and the second inequality follows from (A.7).

Let $\varepsilon_n \equiv ((\log n)^2 m_n^{1/2} \zeta_n^L)^{-1}$. Since $\mathcal{X}$ is compact, there exists a set of points $\{x_i\}_{i=1}^{K_n} \subseteq \mathcal{X}$ with $K_n = O(\varepsilon_n^{-d_x})$ such that each $x \in \mathcal{X}$ can be matched with some x_i satisfying $\|x - x_i\| \leq \varepsilon_n$. For each of such matched pairs, $\{x, x_i\}$, (A.8) implies

$$\begin{aligned}
&\left| \left(\frac{P(x)^\top}{\sigma_{j,n}(x)} - \frac{P(x_i)^\top}{\sigma_{j,n}(x_i)} \right) Q_n^{-1} n^{-1/2} \sum_{t=1}^n P(X_t)\, u_{j,t} \right| \\
&\quad \leq K \zeta_n^L \|x - x_i\| \left\| Q_n^{-1} n^{-1/2} \sum_{t=1}^n P(X_t)\, u_{j,t} \right\| \\
&\quad \leq K (\log n)^{-2} m_n^{-1/2} \left\| Q_n^{-1} n^{-1/2} \sum_{t=1}^n P(X_t)\, u_{j,t} \right\|. \qquad \text{(A.9)}
\end{aligned}$$

In addition, we associate with the set V_n a finite subset $\widetilde{V}_n \subseteq \{1, \ldots, J\} \times \mathcal{X}$ defined as

$$\widetilde{V}_n = \{(j, \tilde{x}) : \tilde{x} \in \{x_i\}_{1 \leq i \leq K_n} \text{ and } \|\tilde{x} - x\| \leq \varepsilon_n \text{ for some } x \text{ such that } (j, x) \in V_n\}.$$

By (A.9), we deduce

$$\begin{aligned}
&\left| \sup_{(j,x) \in V_n} \frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)} n^{-1/2} \sum_{t=1}^n P(X_t)\, u_{j,t} - \max_{(j,x) \in \widetilde{V}_n} \frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)} n^{-1/2} \sum_{t=1}^n P(X_t)\, u_{j,t} \right| \\
&\quad \leq K (\log n)^{-2} m_n^{-1/2} \max_{1 \leq j \leq J} \left\| Q_n^{-1} n^{-1/2} \sum_{t=1}^n P(X_t)\, u_{j,t} \right\| = o_p\left((\log n)^{-1} \right), \qquad \text{(A.10)}
\end{aligned}$$

where the $o_p(\cdot)$ statement follows from $\|Q_n^{-1} n^{-1/2} \sum_{t=1}^n P(X_t)\, u_{j,t}\| = O_p(m_n^{1/2})$.

Define L_n as the cardinality of $\widetilde{V}_n$. Consider a generic Jm_n-vector $z = \left(z_1^\top, \ldots, z_J^\top\right)^\top$ with each z_j component being m_n-dimensional. It is easy to see that we can find Jm_n-vectors α_l, $1 \leq l \leq L_n$, such that for all z,

$$\max_{1 \leq l \leq L_n} \alpha_l^\top z = \max_{(j,x) \in \widetilde{V}_n} \frac{P(x)^\top Q_n^{-1} z_j}{\sigma_{j,n}(x)},$$

and

$$\sup_{1\leq l\leq L_n}\|\alpha_l\|\leq\sup_{(j,x)\in\mathcal{V}}\left\|\frac{P(x)Q_n^{-1}}{\sigma_{j,n}(x)}\right\|\leq K.$$

Note that $L_n=O(K_n)=O(\varepsilon_n^{-d_x})$ satisfies the requirement in Assumption 3, which implies the existence of a random sequence $\widetilde{U}_n$ satisfying $\widetilde{U}_n\overset{d}{=}\max_{1\leq l\leq L_n}\alpha_l^\top\widetilde{N}_n$ and

$$\left|\max_{1\leq l\leq L_n}\alpha_l^\top\left(n^{-1/2}\sum_{t=1}^n u_t\otimes P(X_t)\right)-\widetilde{U}_n\right|=o_p\left((\log n)^{-1}\right). \tag{A.11}$$

By the definition of α_l, we can rewrite (A.11) as

$$\left|\max_{(j,x)\in\widetilde{V}_n}\frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)}n^{-1/2}\sum_{t=1}^n P(X_t)u_{j,t}-\widetilde{U}_n\right|=o_p\left((\log n)^{-1}\right), \tag{A.12}$$

and also note that

$$\widetilde{U}_n\overset{d}{=}\max_{(j,x)\in\widetilde{V}_n}\frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)}\widetilde{N}_{j,n}. \tag{A.13}$$

Following the same argument leading to (A.10), we can also show that

$$\left|\sup_{(j,x)\in V_n}\frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)}\widetilde{N}_{j,n}-\max_{(j,x)\in\widetilde{V}_n}\frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)}\widetilde{N}_{j,n}\right|=o_p\left((\log n)^{-1}\right). \tag{A.14}$$

By (A.13), (A.14), and Lemma 9 in Chernozhukov, Lee, and Rosen (2013), there exists another random sequence $\widetilde{U}_{1,n}$ such that

$$\widetilde{U}_{1,n}\overset{d}{=}\sup_{(j,x)\in V_n}\frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)}\widetilde{N}_{j,n},\quad \widetilde{U}_n-\widetilde{U}_{1,n}=o_p\left((\log n)^{-1}\right). \tag{A.15}$$

From (A.10), (A.12), and (A.15), we further deduce

$$\left|\sup_{(j,x)\in V_n}\frac{P(x)^\top Q_n^{-1}}{\sigma_{j,n}(x)}n^{-1/2}\sum_{t=1}^n P(X_t)u_{j,t}-\widetilde{U}_{1,n}\right|=o_p\left((\log n)^{-1}\right). \tag{A.16}$$

The assertion of the lemma concerning $\widetilde{U}_{1,n}$ then follows from (A.1) and (A.16).

Applying the same arguments with $(\alpha_l)_{1\leq l\leq L_n}$ replaced by $(-\alpha_l)_{1\leq l\leq L_n}$ and $(\alpha_l,-\alpha_l)_{1\leq l\leq L_n}$, we can prove the other two assertions of the lemma, respectively. □

Proof of Proposition 1. We prove this proposition by adapting Chernozhukov, Lee, and Rosen's (2013) proof of their Theorem 2. The main change is to use the three types of coupling results in Theorem A1 in place of Chernozhukov, Lee, and Rosen's Condition C.2 for coupling the entire t-statistic process. To avoid repetition, we relegate the (somewhat tedious) details to the Supplemental Appendix. □

A.2 Proofs of Theorem 1 and Corollary 1

The proof consists of several steps. In the first step (see Lemma A1), we approximate the sequence $\max_{1\leq l\leq L_n}\alpha_l^\top S_n^*$ with $\max_{1\leq l\leq L_n}\alpha_l^\top S_n^+$, where S_n^+ is constructed as a martingale with its predictable quadratic covariation being equal to the deterministic matrix

$$\Sigma_n^* \equiv \sum_{t=1}^{n}\mathbb{E}\left[V_{n,t}^*\right].$$

In the second step (see Lemma A2), we establish a coupling for $\max_{1\leq l\leq L_n}\alpha_l^\top S_n^+$ using Lindeberg's method and Strassen's theorem, for which the fact that Σ_n^* is deterministic is crucial. These approximation results can then be used to construct the coupling for the original $\max_{1\leq l\leq L_n}\alpha_l^\top S_n$ statistic.

We start with constructing the aforementioned martingale S_n^+. The construction is based on the same scheme as in Li and Liao (2020), which we recall as follows. Consider the following stopping time:

$$\tau_n \equiv \max\left\{t\in\{1,\ldots,n\} : \Sigma_n^* - \sum_{s=1}^{t}V_{n,s}^* \text{ is positive semi-definite}\right\},$$

with the convention that $\max\emptyset = 0$. Note that τ_n is a stopping time because $V_{n,t}^*$ is $\mathcal{F}_{n,t-1}$-measurable for each t and Σ_n^* is nonrandom. The matrix

$$\Xi_n \equiv \begin{cases} \Sigma_n^* & \text{when } \tau_n = 0, \\ \Sigma_n^* - \sum_{t=1}^{\tau_n}V_{n,s}^* & \text{when } \tau_n \geq 1, \end{cases}$$

is positive semi-definite by construction.

Let k_n be a sequence of integers such that $k_n\to\infty$ and let $(\eta_{n,t})_{n+1\leq t\leq n+k_n}$ be independent M_n-dimensional standard normal vectors. We construct another martingale difference array $(Z_{n,t},\mathcal{H}_{n,t})_{1\leq t\leq n+k_n}$ as follows:

$$Z_{n,t} \equiv \begin{cases} n^{-1/2}X_{n,t}^*1_{\{t\leq\tau_n\}} & \text{when } 1\leq t\leq n, \\ k_n^{-1/2}\Xi_n^{1/2}\eta_{n,t} & \text{when } n+1\leq t\leq n+k_n, \end{cases}$$

and the filtration is given by

$$\mathcal{H}_{n,t} \equiv \begin{cases} \mathcal{F}_{n,t} & \text{when } 1\leq t\leq n, \\ \mathcal{F}_{n,n}\vee\sigma(\eta_{n,s} : s\leq t) & \text{when } n+1\leq t\leq n+k_n. \end{cases}$$

Since τ_n is a stopping time, it is easy to verify that $(Z_{n,t}, \mathcal{H}_{n,t})_{1\le t\le n+k_n}$ indeed forms a martingale difference array. We denote

$$V_{n,t}^{+} \equiv \mathbb{E}\left[Z_{n,t} Z_{n,t}^{\top} \middle| \mathcal{H}_{n,t-1} \right] \tag{A.17}$$

and set

$$S_n^{+} \equiv \sum_{t=1}^{n+k_n} Z_{n,t}. \tag{A.18}$$

Note that the predictable quadratic covariation matrix of S_n^{+} is exactly Σ_n^{*}, that is,

$$\sum_{t=1}^{n+k_n} V_{n,t}^{+} = \sum_{t=1}^{\tau_n} V_{n,t}^{*} + \Xi_n = \Sigma_n^{*}. \tag{A.19}$$

The approximation error between $\max_{1\le l\le L_n} \alpha_l^{\top} S_n^{*}$ and $\max_{1\le l\le L_n} \alpha_l^{\top} S_n^{+}$ is quantified by Lemma A1, below. We recall from Theorem 1 that $B_{2,n}$ is defined as

$$B_{2,n} \equiv \min\left\{ L_n^{1/q} \max_{l,t} \|\alpha_l^{\top} X_{n,t}^{*}\|_q, M_n^{1/2} \right\} + (\log L_n)^{1/2}.$$

Lemma A1. *Under Assumption 4,* $\left|\max_{1\le l\le L_n} \alpha_l^{\top} S_n^{*} - \max_{1\le l\le L_n} \alpha_l^{\top} S_n^{+}\right| = O_p(B_{2,n} r_n^{1/2})$.

Proof. See the Supplemental Appendix. □

In the next step, we construct a sequence of Gaussian random vectors $\widetilde{S}_n^{*} \sim \mathcal{N}(0, \Sigma_n^{*})$ such that the distribution of $\max_{1\le l\le L_n} \alpha^{\top} \widetilde{S}_n^{*}$ is "close" to that of $\max_{1\le l\le L_n} \alpha_l^{\top} S_n^{+}$ in the sense stated by Lemma A2 below. Specifically, let $(\zeta_{n,t})_{1\le t\le n+k_n}$ be independent M_n-dimensional standard normal vectors that are also independent of $\mathcal{H}_{n,n+k_n}$, and then set

$$\widetilde{S}_n^{*} \equiv \sum_{t=1}^{n+k_n} \tilde{\zeta}_{n,t}, \quad \text{where} \quad \tilde{\zeta}_{n,t} \equiv (V_{n,t}^{+})^{1/2} \zeta_{n,t}.$$

By (A.19), $\widetilde{S}_n^{*} \sim \mathcal{N}(0, \Sigma_n^{*})$. The next lemma quantifies the difference between the distributions of $\max_{1\le l\le L_n} \alpha_l^{\top} S_n^{+}$ and $\max_{1\le l\le L_n} \alpha_l^{\top} \widetilde{S}_n^{*}$. Below, for any Borel subset $A \subseteq \mathbb{R}$ and any constant $\delta > 0$, we denote the δ-enlargement of A by A^{δ}, that is,

$$A^{\delta} \equiv \left\{ x \in \mathbb{R} : \inf_{y\in A} \|x - y\| \le \delta \right\}.$$

We remind the reader that the $B_{1,n}$ sequence is defined in Theorem 1.

Lemma A2. *Suppose that Assumption 4 holds and* $k_n \ge (\log L_n)^3 B_{1,n}^{-2}$. *Then, for each* $C > 5$,

$$\mathbb{P}\left(\max_{1\le l\le L_n} \alpha_l^{\top} S_n^{+} \in A \right) \le \mathbb{P}\left(\max_{1\le l\le L_n} \alpha^{\top} \widetilde{S}_n^{*} \in A^{C B_{1,n}^{1/3} (\log L_n)^{2/3}} \right) + \epsilon_n(C),$$

where $\epsilon_n(C)$ *is a real sequence satisfying* $\sup_{n\ge 1} \epsilon_n(C) \to 0$ *as* $C \to \infty$.

PROOF. See the Supplemental Appendix. □

We are now ready to prove Theorem 1 and its corollary.

PROOF OF THEOREM 1. Since $B_{1,n}^{1/3}(\log L_n)^{2/3} = o((\log n)^{-1})$ by assumption, we can find a real sequence $C_n \to \infty$ such that

$$C_n B_{1,n}^{1/3}(\log L_n)^{2/3} = o((\log n)^{-1}).$$

By Lemma A2 (for which the condition on k_n can be trivially verified by taking k_n sufficiently large), we have for each $n \geq 1$,

$$\mathbb{P}\left(\max_{1\leq l\leq L_n} \alpha_l^\top S_n^+ \in A\right) \leq \mathbb{P}\left(\max_{1\leq l\leq L_n} \alpha^\top \widetilde{S}_n^* \in A^{C_n B_{1,n}^{1/3}(\log L_n)^{2/3}}\right) + \epsilon_n(C_n).$$

By Strassen's theorem, for each n, we can construct a random variable $\widetilde{U}_n^*$ such that $\widetilde{U}_n^* \stackrel{d}{=} \max_{1\leq l\leq L_n} \alpha^\top \widetilde{S}_n^*$ and

$$\mathbb{P}\left(\left|\max_{1\leq l\leq L_n} \alpha_l^\top S_n^+ - \widetilde{U}_n^*\right| > o((\log n)^{-1})\right) \leq \epsilon_n(C_n).$$

Since $\epsilon_n(C_n) \to 0$, we deduce from the above estimate that

$$\max_{1\leq l\leq L_n} \alpha_l^\top S_n^+ - \widetilde{U}_n^* = o_p\left((\log n)^{-1}\right).$$

By Lemma A1, we further have

$$\max_{1\leq l\leq L_n} \alpha_l^\top S_n^* - \widetilde{U}_n^* = o_p\left((\log n)^{-1}\right). \tag{A.20}$$

Let $\widetilde{S}_n \equiv (\Sigma_n)^{1/2}(\Sigma_n^*)^{-1/2}\widetilde{S}_n^*$. Note that $\widetilde{S}_n \sim \mathcal{N}(0, \Sigma_n)$. By definition,

$$\alpha_l^\top \widetilde{S}_n - \alpha_l^\top \widetilde{S}_n^* = \alpha_l^\top((\Sigma_n)^{1/2} - (\Sigma_n^*)^{1/2})(\Sigma_n^*)^{-1/2}\widetilde{S}_n^*$$

where $(\Sigma_n^*)^{-1/2}\widetilde{S}_n^*$ is a standard normal random vector. By the triangle inequality,

$$\left|\max_{1\leq l\leq L_n} \alpha_l^\top \widetilde{S}_n - \max_{1\leq l\leq L_n} \alpha_l^\top \widetilde{S}_n^*\right| \leq \max_{1\leq l\leq L_n} \left|\alpha_l^\top \widetilde{S}_n - \alpha_l^\top \widetilde{S}_n^*\right|.$$

With an appeal to the maximum inequality, we deduce

$$\begin{aligned}
\mathbb{E}\left[\max_{1\leq l\leq L_n} \left|\alpha_l^\top \widetilde{S}_n - \alpha_l^\top \widetilde{S}_n^*\right|\right] &\leq K(\log L_n)^{1/2} \max_{1\leq l\leq L_n} \left\|\alpha_l^\top((\Sigma_n)^{1/2} - (\Sigma_n^*)^{1/2})\right\| \\
&\leq K(\log L_n)^{1/2} \|\Sigma_n - \Sigma_n^*\|_S \\
&= O((\log L_n)^{1/2} \bar{c}_n M_n^{1/2} n^{-1/2}),
\end{aligned}$$

where the last line is by (A.79) in the Supplemental Appendix of Li and Liao (2020). Combining the above estimates, we have

$$\left| \max_{1\leq l\leq L_n} \alpha_l^\top \widetilde{S}_n - \max_{1\leq l\leq L_n} \alpha_l^\top \widetilde{S}_n^* \right| = O_p\left((\log L_n)^{1/2} \bar{c}_n M_n^{1/2} n^{-1/2} \right) = o_p\left((\log n)^{-1} \right).$$

By this estimate and Lemma 9 in Chernozhukov, Lee, and Rosen (2013), we can construct another random sequence $\widetilde{U}_n$ such that

$$\widetilde{U}_n \stackrel{d}{=} \max_{1\leq l\leq L_n} \alpha_l^\top \widetilde{S}_n, \quad \widetilde{U}_n - \widetilde{U}_n^* = o_p\left((\log n)^{-1} \right). \tag{A.21}$$

The assertion of the theorem then follows from Assumption 4(i), (A.20), and (A.21). □

Proof of Corollary 1. The proof is done by verifying the conditions of Theorem 1. Since $V_{n,t}^*$ is deterministic, we can set

$$r_n = 0. \tag{A.22}$$

By the boundedness of (α_l) and the Cauchy–Schwarz inequality, $\left|\alpha_l^\top X_{n,t}^*\right| \leq K \left\|X_{n,t}^*\right\|$. Hence,

$$\begin{aligned} \max_{1\leq l\leq L_n} \left(\mathbb{E}\left[(\alpha_l^\top X_{n,t}^*)^2 | \mathcal{F}_{n,t-1} \right] \right)^{3/2} &\leq K \left(\mathbb{E}\left[\left\|X_{n,t}^*\right\|^2 | \mathcal{F}_{n,t-1} \right] \right)^{3/2} \\ &\leq K \mathbb{E}\left[\left\|X_{n,t}^*\right\|^3 | \mathcal{F}_{n,t-1} \right], \end{aligned}$$

where the second inequality is by Jensen's inequality. It is then easy to see that

$$\begin{aligned} B_{1,n} &\equiv n^{-3/2} \sum_{t=1}^n \mathbb{E}\left[(\log L_n)^{3/2} \max_{1\leq l\leq L_n} \left(\mathbb{E}\left[(\alpha_l^\top X_{n,t}^*)^2 | \mathcal{F}_{n,t-1} \right] \right)^{3/2} + \max_{1\leq l\leq L_n} |\alpha_l^\top X_{n,t}^*|^3 \right] \\ &\leq K n^{-3/2} \sum_{t=1}^n \left((\log L_n)^{3/2} + 1 \right) \mathbb{E}\left[\left\|X_{n,t}^*\right\|^3 \right] \leq K \bar{c}_n^3 (\log L_n)^{3/2} n^{-1/2} M_n^{3/2}. \end{aligned}$$

Therefore,

$$B_{1,n}^{1/3} (\log M_n)^{2/3} \leq K \bar{c}_n (\log L_n)^{1/2} (\log M_n)^{2/3} M_n^{1/2} n^{-1/6}.$$

Since M_n and L_n both have polynomial growth in n, $\log L_n = O(\log n)$ and $\log M_n = O(\log n)$. Therefore,

$$B_{1,n}^{1/3} (\log M_n)^{2/3} = o\left((\log n)^2 n^{\frac{a}{2} - \frac{1}{6}} \right) = o((\log n)^{-1}). \tag{A.23}$$

By a similar argument, it is easy to see that

$$\bar{c}_n (\log L_n)^{1/2} M_n^{1/2} n^{-1/2} = o((\log n)^{-1}). \tag{A.24}$$

With (A.22), (A.23), and (A.24), we readily verify the conditions of Theorem 1, which finishes the proof of this corollary. □