Asymptotic Inference about Predictive Accuracy using High Frequency Data*

Jia Li
Department of Economics
Duke University

Andrew J. Patton
Department of Economics
Duke University

This Version: July 6, 2013

Abstract

This paper provides a general framework that enables many existing inference methods for predictive accuracy to be used in applications that involve forecasts of latent target variables. Such applications include the forecasting of volatility, correlation, beta, quadratic variation, jump variation, and other functionals of an underlying continuous-time process. We provide primitive conditions under which a "negligibility" result holds, and thus the asymptotic size of standard predictive accuracy tests, implemented using high-frequency proxies for the latent variable, is controlled. An extensive simulation study verifies that the asymptotic results apply in a range of empirically relevant applications, and an empirical application to correlation forecasting is presented.

Keywords: Forecast evaluation, realized variance, volatility, jumps, semimartingale.
JEL Codes: C53, C22, C58, C52, C32.

*Financial support from the NSF under grant SES-1227448 (Li) is gratefully acknowledged.

Contact address: Department of Economics, Duke University, 213 Social Sciences Building, Box 90097, Durham NC 27708-0097, USA. Email: jl410@duke.edu and andrew.patton@duke.edu.

1 Introduction

A central problem in times series analysis is the forecasting of economic variables, and in financial applications the variables to be forecast are often risk measures, such as volatility, beta, correlation, or jump variation. Since the seminal work of Engle (1982), numerous models have been proposed to forecast risk measures, and these forecasts are of fundamental importance in financial decisions. The problem of evaluating the performance of these forecasts is complicated by the fact that many risk measures, although well-defined in models, are not directly observable. A large literature has evolved presenting methods for inference for forecast accuracy, however existing work typically relies on the observability of the forecast target; see Diebold and Mariano (1995), West (1996), White (2000), Giacomini and White (2006), McCracken (2007), Romano and Wolf (2005), and Hansen, Lunde, and Nason (2011), as well as West (2006) for a review. The goal of the current paper is to extend the applicability of the aforementioned methods to settings with an unobservable forecast target variable.

Our proposal is to implement the standard forecast evaluation methods, such as those mentioned above, with the unobservable target variable replaced by a proxy computed using high-frequency (intraday) data. Competing forecasts are evaluated with respect to the proxy by using existing inference methods proposed in the above papers. *Prima facie*, such inference is not of direct economic interest, in that a good forecast for the proxy may not be a good forecast of the latent target variable. The gap, formally speaking, arises from the fact that hypotheses concerning the proxy are not the same as those concerning the true target variable. We fill this gap by providing high-level conditions that lead to a "negligibility" result, which shows that the asymptotic level and power properties of the existing inference methods are valid not only under the "proxy hypotheses," but also under the "true hypotheses." The theoretical results are supported by an extensive and realistically calibrated Monte Carlo study.

The high-level assumptions underlying our theory broadly involve two conditions. The first condition imposes an abstract structure on the inference methods with an observable target variable, which enables us to cover many predictive accuracy methods proposed in the literature as special cases, including almost all of the papers cited above. The second condition concerns the approximation accuracy of the proxy relative to the latent target variable, and we provide primitive conditions for general classes of high-frequency based estimators of volatility and jump functionals, which cover almost all existing estimators as special cases, such as realized (co)variation, truncated (co)variation, bipower variation, realized correlation, realized beta, jump power variation, realized semivariance, realized Laplace transform, realized skewness and kurtosis.

The main contribution of the current paper is methodological: we provide a simple but general framework for studying the problem of testing for predictive ability with a latent target variable.

Our results provide ex-post justification for existing empirical results on forecast evaluation using high-frequency proxies, and can readily be applied to promote further studies on a wide spectrum of risk measures and high-frequency proxies using a variety of evaluation methods. In obtaining our main result we make two other contributions. The first is primarily pedagogical: we present a simple unifying framework for considering much of the extant literature on forecast evaluation, including Diebold and Mariano (1995), West (1996), McCracken (2000), White (2000), Giacomini and White (2006), and McCracken (2007), which also reveals avenues for further extension to the framework proposed here. The second auxiliary contribution is technical: in the process of verifying our high-level assumptions on proxy accuracy, we provide results on the rate of convergence for a comprehensive collection of high-frequency based estimators for general multivariate Itô semimartingale models. Such results may be used in other applications involving high-frequency proxies, such as the estimation and specification problems considered by Corradi and Distaso (2006), Corradi, Distaso, and Swanson (2009, 2011) and Todorov and Tauchen (2012b).

We illustrate our approach in an application involving competing forecasts of the conditional correlation between stock returns. We consider four forecasting methods, starting with the popular "dynamic conditional correlation" (DCC) model of Engle (2002). We then extend this model to include an asymmetric term, as in Cappiello, Engle, and Sheppard (2006), which allows correlations to rise more following joint negative shocks than other shocks, and to include the lagged realized correlation matrix, which enables the model to exploit higher frequency data, in the spirit of Noureldin, Shephard, and Sheppard (2012). We find evidence, across a range of correlation proxies, that including high frequency information in the forecast model leads to out-of-sample gains in accuracy, while the inclusion of an asymmetric term does not lead to such gains.

The existing literature includes some work on forecast evaluation for latent target variables using proxy variables. In their seminal work, Andersen and Bollerslev (1998) advocate using the realized variance as a proxy for evaluating volatility forecast models; also see Andersen, Bollerslev, Diebold, and Labys (2003), Andersen, Bollerslev, and Meddahi (2005) and Andersen, Bollerslev, Christoffersen, and Diebold (2006). A theoretical justification for this approach was proposed by Hansen and Lunde (2006) and Patton (2011), based on the availability of conditionally unbiased proxies. Their unbiasedness condition must hold in finite samples, which is generally hard to verify except for specially designed examples. In contrast, our framework uses an asymptotic argument and is applicable for most known high-frequency based estimators, as shown in Section 3.

The current paper is also related to the large and growing literature on high-frequency econometrics (see Jacod and Protter (2012)). In the process of verifying our high-level assumptions on the approximation accuracy of the proxies, we provide rates of convergence for general classes of high-frequency based estimators. These results are related to, but different from, the fill-in asymptotic result used by Jacod and Protter (2012), among others. Indeed, we consider a large-T

asymptotic setting with the mesh of the possibly irregular sampling grids of high-frequency data going to zero in "later" sample periods. We refer to our asymptotic setting as "eventually fill-in asymptotics." Moreover, we also consider (oft-neglected) situations in which the asymptotic distribution of the high frequency estimator is unavailable, such as realized skewness, bipower variation and semivariance in the presence of jumps, and truncation-based estimators in cases with "active jumps." Further technical discussion of the literature is presented in Section 3.6.

The paper is organized as follows. Section 2 presents the main theory. Section 3 verifies high-level assumptions on the proxy under primitive conditions. Section 4 provides extensions of some popular forecast evaluation methods that do not fit directly into our baseline framework. Monte Carlo results and an empirical application are in Sections 5 and 6, respectively. All proofs are in the Supplemental Material to this paper, which also contains some additional simulation results.

Notation

All limits below are for $T \to \infty$. We use $\xrightarrow{\mathbb{P}}$ to denote convergence in probability and $\xrightarrow{d}$ to denote convergence in distribution. All vectors are column vectors. For any matrix A, we denote its transpose by $A^\intercal$ and its (i,j) component by A_{ij}. The (i,j) component of a matrix-valued stochastic process A_t is denoted by $A_{ij,t}$. We write (a,b) in place of $(a^\intercal, b^\intercal)^\intercal$. The jth component of a vector x is denoted by x_j. For $x, y \in \mathbb{R}^q$, $q \geq 1$, we write $x \leq y$ if and only if $x_j \leq y_j$ for any $j \in \{1, \ldots, q\}$. For a generic variable X taking values in a finite-dimensional space, we use κ_X to denote its dimensionality; the letter κ is reserved for such use. The time index t is interpreted in continuous time. For simplicity, we refer to the time unit as a "day" while it can also be a week, a month, etc.; the words "daily", "intraday", "intradaily" should be interpreted accordingly. We use $\|\cdot\|$ to denote the Euclidean norm of a vector, where a matrix is identified as its vectorized version. For each $p \geq 1$, $\|\cdot\|_p$ denotes the L_p norm. We use $\circ$ to denote the Hadamard product between two identically sized matrices, which is computed simply by element-by-element multiplication. The notation $\otimes$ stands for the Kronecker product. For two sequences of strictly positive real numbers a_t and b_t, $t \geq 1$, we write $a_t \asymp b_t$ if and only if the sequences a_t/b_t and b_t/a_t are both bounded.

2 The main theory

This section presents the main theoretical results of the paper based on high level conditions. In Section 2.1 we link existing tests of predictive ability into a unified framework, and in Section 2.2 we consider the extension of these tests to handle latent target variables and present the main theorem (Theorem 2.1). Primitive conditions for the main theorem are presented in the next section.

2.1 Testing predictive accuracy with an observable target

We start with the basic problem with an *observable* forecast target. Let Y_t be the time series to be forecast, taking values in $\mathcal{Y} \subseteq \mathbb{R}^{\kappa_Y}$. At time t, the forecaster uses data $\mathcal{D}_t \equiv \{D_s : 1 \leq s \leq t\}$ to form a forecast of $Y_{t+\tau}$, where the horizon $\tau \geq 1$ is fixed throughout the paper. We consider $\bar{k}$ competing sequences of forecasts of $Y_{t+\tau}$, collected by $F_{t+\tau} = (F_{1,t+\tau}, \ldots, F_{\bar{k},t+\tau})$. In practice, $F_{t+\tau}$ is often constructed from forecast models involving some parameter β which is typically finite-dimensional but may be infinite-dimensional if nonparametric techniques are involved. We write $F_{t+\tau}(\beta)$ to emphasize such dependence, and refer to the function $F_{t+\tau}(\cdot)$ as the forecast model. Let $\hat{\beta}_t$ be an estimator of β using (possibly a subset of) the dataset $\mathcal{D}_t$ and β^* be its "population" analogue.[1]

We sometimes need to distinguish two types of forecasts: the actual forecast $F_{t+\tau} = F_{t+\tau}(\hat{\beta}_t)$ and the population forecast $F_{t+\tau}(\beta^*)$. This distinction is useful when a researcher is interested in using the actual forecast $F_{t+\tau}$ to make inference concerning $F_{t+\tau}(\beta^*)$, that is, an inference concerning the forecast model (see, e.g., West (1996)). If, on the other hand, the researcher is interested in assessing the performance of the actual forecasts in $F_{t+\tau}$, she can set β^* to be empty and treat the actual forecast as an observable sequence (see, e.g., Diebold and Mariano (1995) and Giacomini and White (2006)). Therefore, an inference framework concerning $F_{t+\tau}(\beta^*)$ can also be used to make inference for the actual forecasts; we hence adopt this general setting in our framework.[2]

Given the target $Y_{t+\tau}$, the performance of the competing forecasts in $F_{t+\tau}$ is measured by $f_{t+\tau} \equiv f(Y_{t+\tau}, F_{t+\tau}(\hat{\beta}_t))$, where $f(\cdot)$ is a known measurable $\mathbb{R}^{\kappa_f}$-valued function. Typically, $f(\cdot)$ is the loss differential between competing forecasts. We also denote $f^*_{t+\tau} \equiv f(Y_{t+\tau}, F_{t+\tau}(\beta^*))$ and set

$$\bar{f}_T \equiv P^{-1} \sum_{t=R}^{T} f_{t+\tau}, \quad \bar{f}^*_T \equiv P^{-1} \sum_{t=R}^{T} f^*_{t+\tau}, \tag{2.1}$$

where T is the size of the full sample, $P = T - R + 1$ is the size of the prediction sample and R is the size of the estimation sample.[3] In the sequel, we always assume $P \asymp T$ as $T \to \infty$ without further mention, while R may be fixed or goes to ∞, depending on the application.

Our baseline theory concerns two classical testing problems for forecast evaluation: testing for equal predictive ability (one-sided or two-sided) and testing for superior predictive ability.

[1] If the forecast model is correctly specified, β^* is the true parameter of the model. In general, β^* is considered as the pseudo-true parameter.

[2] The "generality" here should only be interpreted in a notational, instead of an econometric, sense, as the econometric scope of Diebold and Mariano (1995) and Giacomini and White (2006) is very different from that of West (1996). See Giacomini and White (2006) and Diebold (2012) for more discussion.

[3] The notations P_T and R_T may be used in place of P and R. We follow the literature and suppress the dependence on T. The estimation and prediction samples are often called the in-sample and (pseudo-) out-of-sample periods.

Formally, we consider the following hypotheses: for some user-specified constant $\chi \in \mathbb{R}^{\kappa_f}$,

$$\begin{array}{c}\text{Proxy Equal}\\ \text{Predictive Ability}\\ \text{(PEPA)}\end{array} \begin{cases} & H_0 : \mathbb{E}\left[f^*_{t+\tau}\right] = \chi \text{ for all } t \geq 1, \\ \text{vs.} & H_{1a} : \liminf_{T\to\infty} \mathbb{E}[\bar{f}^*_{j,T}] > \chi_j \text{ for some } j \in \{1, \ldots, \kappa_f\}, \\ \text{or} & H_{2a} : \liminf_{T\to\infty} \|\mathbb{E}[\bar{f}^*_T] - \chi\| > 0, \end{cases} \tag{2.2}$$

$$\begin{array}{c}\text{Proxy Superior}\\ \text{Predictive Ability}\\ \text{(PSPA)}\end{array} \begin{cases} & H_0 : \mathbb{E}[f^*_{t+\tau}] \leq \chi \text{ for all } t \geq 1, \\ \text{vs.} & H_a : \liminf_{T\to\infty} \mathbb{E}[\bar{f}^*_{j,T}] > \chi_j \text{ for some } j \in \{1, \ldots, \kappa_f\}, \end{cases} \tag{2.3}$$

where H_{1a} (resp. H_{2a}) in (2.2) is the one-sided (resp. two-sided) alternative. In practice, χ is often set to be zero.[4] In Section 2.2 below, $Y_{t+\tau}$ plays the role of a proxy for the latent true forecast target, which explains the qualifier "proxy" in the labels of the hypotheses above. These hypotheses allow for data heterogeneity and are cast in the same fashion as in Giacomini and White (2006). Under (mean) stationarity, these hypotheses coincide with those considered by Diebold and Mariano (1995), West (1996) and White (2000), among others. We note that, by setting the function $f(\cdot)$ properly, the hypotheses in (2.2) can also be used to test for forecast encompassing and forecast unbiasedness.[5]

We consider a test statistic of the form

$$\varphi_T \equiv \varphi(a_T(\bar{f}_T - \chi), a'_T S_T) \tag{2.4}$$

for some measurable function $\varphi : \mathbb{R}^{\kappa_f} \times \mathcal{S} \mapsto \mathbb{R}$, where $a_T \to \infty$ and a'_T are known deterministic sequences, and S_T is a sequence of $\mathcal{S}$-valued estimators that is mainly used for studentization.[6] In almost all cases, $a_T = P^{1/2}$ and $a'_T \equiv 1$; recall that P increases with T. An exception is given by Example 2.4 below. In many applications, S_T plays the role of an estimator of some asymptotic variance, which may or may not be consistent (see Example 2.2 below); $\mathcal{S}$ is then the space of positive definite matrices. Further concrete examples are given below.

Let $\alpha \in (0, 1)$ be the significance level of a test. We consider a (nonrandomized) test of the form $\phi_T = \mathbf{1}\{\varphi_T > z_{T,1-\alpha}\}$, that is, we reject the null hypothesis when the test statistic φ_T is greater than some critical value $z_{T,1-\alpha}$. We now introduce some high-level assumptions on the test statistic and the critical value for conducting tests based on PEPA and PSPA.

Assumption A1: $(a_T(\bar{f}_T - \mathbb{E}[\bar{f}^*_T]), a'_T S_T) \xrightarrow{d} (\xi, S)$ for some deterministic sequence $a_T \to \infty$ and a'_T, and random variables (ξ, S). Here, (a_T, a'_T) may be chosen differently under the null and the alternative hypotheses, but φ_T is invariant to such choice.

[4] Allowing χ to be nonzero incurs no additional cost in our deriviations. This flexibility is useful in the design of Monte Carlo experiment that examines the finite-sample performance of the asymptotic theory below. See Section 5 for details.

[5] See p. 109 in West (2006).

[6] The space $\mathcal{S}$ changes across applications, but is always implicitly assumed to be a Polish space.

Assumption A1 covers many existing methods as special cases. We discuss a list of examples below for concreteness.

Example 2.1: Giacomini and White (2006) consider tests for equal predictive ability between two sequences of actual forecasts, or "forecast methods" in their terminology, assuming R fixed. In this case, $f(Y_t,(F_{1,t},F_{2,t})) = L(Y_t,F_{1,t}) - L(Y_t,F_{2,t})$ for some loss function $L(\cdot,\cdot)$. Moreover, one can set β^* to be empty and treat each actual forecast as an observed sequence; in particular, $f_{t+\tau} = f^*_{t+\tau}$ and $\bar{f}_T = \bar{f}^*_T$. Using a CLT for heterogeneous weakly dependent data, one can take $a_T = P^{1/2}$ and verify $a_T(\bar{f}_T - \mathbb{E}[\bar{f}_T]) \xrightarrow{d} \xi$, where ξ is centered Gaussian with its long-run variance denoted by Σ. We then set $S = \Sigma$ and $a'_T \equiv 1$, and let S_T be a HAC estimator of S (Newey and West (1987), Andrews (1991)). Assumption A1 then follows from Slutsky's lemma. Diebold and Mariano (1995) intentionally treat the actual forecasts as primitives without introducing the forecast model (and hence β^*); their setting is also covered by Assumption A1 by the same reasoning.

Example 2.2: Consider the same setting as in Example 2.1, but let S_T be an inconsistent long-run variance estimator of Σ as considered by, for example, Kiefer and Vogelsang (2005). Using their theory, we verify $(P^{1/2}(\bar{f}_T - \mathbb{E}[\bar{f}_T]), S_T) \xrightarrow{d} (\xi, S)$, where S is a (nondegenerate) random matrix and the joint distribution of ξ and S is known, up to the unknown parameter Σ, but is nonstandard.

Example 2.3: West (1996) considers inference on nonnested forecast models in a setting with $R \to \infty$. West's Theorem 4.1 shows that $P^{1/2}(\bar{f}_T - \mathbb{E}[\bar{f}^*_T]) \xrightarrow{d} \xi$, where ξ is centered Gaussian with its variance-covariance matrix denoted here by S, which captures both the sampling variability of the forecast error and the discrepancy between $\hat{\beta}_t$ and β^*. We can set S_T to be the consistent estimator of S as proposed in West's comment 6 to Theorem 4.1. Assumption A1 is then verified by using Slutsky's lemma for $a_T = P^{1/2}$and $a'_T \equiv 1$. West's theory relies on the differentiability of the function $f(\cdot)$ with respect to β and concerns $\hat{\beta}_t$ in the recursive scheme. Similar results allowing for a nondifferentiable $f(\cdot)$ function can be found in McCracken (2000); Assumption A1 can be verified similarly in this more general setting.

Example 2.4: McCracken (2007) considers inference on nested forecast models allowing for recursive, rolling and fixed estimation schemes, all with $R \to \infty$. The evaluation measure $f_{t+\tau}$ is the difference between the quadratic losses of the nesting and the nested models. For his OOS-t test, McCracken proposes using a normalizing factor $\widehat{\Omega}_T = P^{-1}\sum_{t=R}^{T}(f_{t+\tau} - \bar{f}_T)^2$ and consider the test statistic $\varphi_T \equiv \varphi(P\bar{f}_T, P\widehat{\Omega}_T)$, where $\varphi(u,s) = u/\sqrt{s}$. Implicitly in his proof of Theorem 3.1, it is shown that under the null hypothesis of equal predictive ability, $(P(\bar{f}_T - \mathbb{E}[\bar{f}^*_T]), P\widehat{\Omega}_T) \xrightarrow{d} (\xi, S)$, where the joint distribution of (ξ, S) is nonstandard and is specified as a function of a multivariate Brownian motion. Assumption A1 is verified with $a_T = P$, $a'_T \equiv P$ and $S_T = \widehat{\Omega}_T$. The nonstandard rate arises as a result of the degeneracy between correctly specified nesting models. Under the alternative hypothesis, it can be shown that Assumption A1 holds for $a_T = P^{1/2}$ and $a'_T \equiv 1$, as in West (1996). Clearly, the OOS-t test statistic is invariant to the change of (a_T, a'_T), that is, $\varphi_T = \varphi(P^{1/2}\bar{f}_T, \widehat{\Omega}_T)$ holds. Assumption A1 can also be verified for various (partial) extensions of McCracken (2007); see, for example, Inoue and Kilian (2004), Clark and McCracken (2005) and Hansen and Timmermann (2012).

Example 2.5: White (2000) considers a setting similar to West (1996), with an emphasis on considering a large number of competing forecasts, but uses a test statistic without studentization. Assumption A1 is verified similarly as in Example 2.3, but with S_T and S being empty.

Assumption A2: $\varphi(\cdot,\cdot)$ is continuous almost everywhere under the law of (ξ, S).

Assumption A2 is satisfied by all standard test statistics in this literature: for simple pair-wise forecast comparisons, the test statistic usually takes the form of t-statistic, that is, $\varphi_{\text{t-stat}}(\xi, S) = \xi/\sqrt{S}$. For joint tests it may take the form of a Wald-type statistic, $\varphi_{\text{Wald}}(\xi, S) = \xi^{\intercal} S^{-1}\xi$, or a maximum over individual (possibly studentized) test statistics $\varphi_{\text{Max}}(\xi, S) = \max_i \xi_i$ or $\varphi_{\text{StuMax}}(\xi, S) = \max_i \xi_i/\sqrt{S_i}$.

Assumption A2 imposes continuity on $\varphi(\cdot,\cdot)$ in order to facilitate the use of the continuous mapping theorem for studying the asymptotics of the test statistic φ_T. More specifically, under the null hypothesis of PEPA, which is also the null least favorable to the alternative in PSPA (White (2000), Hansen (2005)), Assumption A1 implies that $(a_T(\bar{f}_T - \chi), a'_T S_T) \xrightarrow{d} (\xi, S)$. By the continuous mapping theorem, Assumption A2 then implies that the asymptotic distribution of φ_T under this null is $\varphi(\xi, S)$. The critical value of a test at nominal level α is given by the $1-\alpha$ quantile of $\varphi(\xi, S)$, on which we impose the following condition.

Assumption A3: The distribution function of $\varphi(\xi, S)$ is continuous at its $1-\alpha$ quantile $z_{1-\alpha}$. Moreover, the sequence $z_{T,1-\alpha}$ of critical values satisfies $z_{T,1-\alpha} \xrightarrow{\mathbb{P}} z_{1-\alpha}$.

The first condition in Assumption A3 is very mild. Assumption A3 is mainly concerned with the availability of the consistent estimator of the $1-\alpha$ quantile $z_{1-\alpha}$. Examples are given below.

Example 2.6: In many cases, the limit distribution of φ_T under the null of PEPA is standard normal or chi-square with some known number of degrees of freedom. Examples include tests considered by Diebold and Mariano (1995), West (1996) and Giacomini and White (2006). In the setting of Example 2.2 or 2.4, φ_T is a t-statistic or Wald-type statistic, with an asymptotic distribution that is nonstandard but pivotal, with quantiles tabulated in the original papers.[7]

[7] One caveat is that the asymptotic pivotalness of the OOS-t and OOS-F statistics in McCracken (2007) are valid under the somewhat restrictive condition that the forecast error forms a conditionally homoskedastic martingale difference sequence. In the presence of conditional heteroskedasticity or serial correlation in the forecast errors, the

Assumption A3 for these examples can be verified by simply taking $z_{T,1-\alpha}$ as the known quantile of the limit distribution.

Example 2.7: White (2000) considers tests for superior predictive ability. Under the null least favorable to the alternative, White's test statistic is not asymptotically pivotal, as it depends on the unknown variance of the limit variable ξ. White suggests computing the critical value via either simulation or the stationary bootstrap (Politis and Romano (1994)), corresponding respectively to his "Monte Carlo reality check" and "bootstrap reality check" methods. In particular, under stationarity, White shows that the bootstrap critical value consistently estimates $z_{1-\alpha}$.[8] Hansen (2005) considers test statistics with studentization and shows the validity of a refined bootstrap critical value, under stationarity. The validity of the stationary bootstrap holds in more general settings allowing for moderate heterogeneity (Gonçalves and White (2002), Gonçalves and de Jong (2003)). We hence conjecture that the bootstrap results of White (2000) and Hansen (2005) can be extended to a setting with moderate heterogeneity, although a formal discussion is beyond the scope of the current paper. In these cases, the simulation- or bootstrap-based critical value can be used as $z_{T,1-\alpha}$ in order to verify Assumption A3.

Finally, we need two alternative sets of assumptions on the test function $\varphi(\cdot,\cdot)$ for one-sided and two-sided tests, respectively.

Assumption B1: For any $s \in \mathcal{S}$, we have (a) $\varphi(u,s) \leq \varphi(u',s)$ whenever $u \leq u'$, where $u, u' \in \mathbb{R}^{\kappa_f}$; (b) $\varphi(u,\tilde{s}) \to \infty$ whenever $u_j \to \infty$ for some $1 \leq j \leq \kappa_f$ and $\tilde{s} \to s$.

Assumption B2: For any $s \in \mathcal{S}$, $\varphi(u,\tilde{s}) \to \infty$ whenever $\|u\| \to \infty$ and $\tilde{s} \to s$.

Assumption B1(a) imposes monotonicity on the test statistic as a function of the evaluation measure, and is used for size control in the PSPA setting. Assumption B1(b) concerns the consistency of the test against the one-sided alternative and is easily verified for commonly used one-sided test statistics, such as $\varphi_{\text{t-stat}}$, φ_{Max} and φ_{StuMax} described in the comment following Assumption A2. Assumption B2 serves a similar purpose for two-sided tests, and is also easily verifiable.

We close this subsection by summarizing the level and power properties of the test ϕ_T.

Proposition 2.1: The following statements hold under Assumptions A1–A3. (a) Under the PEPA setting (2.2), $\mathbb{E}\phi_T \to \alpha$ under H_0. If Assumption B1(b) (resp. B2) holds in addition, we

null distribution generally depends on a nuisance parameter (Clark and McCracken (2005)). Nevertheless, the critical values can be consistently estimated via a bootstrap (Clark and McCracken (2005)) or plug-in method (Hansen and Timmermann (2012)).

[8] White (2000) shows the validity of the bootstrap critical value in a setting where the sampling error in $\hat{\beta}_t$ is asymptotically irrelevant (West (1996), West (2006)). Corradi and Swanson (2007) propose a bootstrap critical value in the general setting of West (1996), without imposing asymptotic irrelavance.

have $\mathbb{E}\phi_T \to 1$ under H_{1a} (resp. H_{2a}). (b) Under the PSPA setting (2.3) and Assumption B1, we have $\limsup_{T\to\infty} \mathbb{E}\phi_T \le \alpha$ under H_0 and $\mathbb{E}\phi_T \to 1$ under H_a.

2.2 Testing predictive accuracy with an unobservable target

We now deviate from the classical setting in Section 2.1. We suppose that the observable series $Y_{t+\tau}$ is not the forecast target of interest, but only a proxy for the true *latent* target series $Y^\dagger_{t+\tau}$. The classical methods for comparing predictive accuracy based on the proxy are statistically valid for the PEPA and PSPA hypotheses. However, these hypotheses are not of immediate economic relevance, because economic agents are, by assumption in this subsection, interested in forecasting the true target $Y^\dagger_{t+\tau}$, rather than its proxy.[9] Formally, we are interested in testing the following "true" hypotheses: for $f^\dagger_{t+\tau} \equiv f(Y^\dagger_{t+\tau}, F_{t+\tau}(\beta^*))$,

$$\begin{array}{c}\text{Equal}\\ \text{Predictive Ability}\\ \text{(EPA)}\end{array} \left\{\begin{array}{ll} & H_0 : \mathbb{E}[f^\dagger_{t+\tau}] = \chi \text{ for all } t \ge 1, \\ \text{vs.} & H_{1a} : \liminf_{T\to\infty} \mathbb{E}[\bar{f}^\dagger_{j,T}] > \chi_j \text{ for some } j \in \{1,\ldots,\kappa_f\}, \\ \text{or} & H_{2a} : \liminf_{T\to\infty} \|\mathbb{E}[\bar{f}^\dagger_T] - \chi\| > 0, \end{array}\right. \tag{2.5}$$

$$\begin{array}{c}\text{Superior}\\ \text{Predictive Ability}\\ \text{(SPA)}\end{array} \left\{\begin{array}{ll} & H_0 : \mathbb{E}[f^\dagger_{t+\tau}] \le \chi \text{ for all } t \ge 1, \\ \text{vs.} & H_a : \liminf_{T\to\infty} \mathbb{E}[\bar{f}^\dagger_{j,T}] > \chi_j \text{ for some } j \in \{1,\ldots,\kappa_f\}. \end{array}\right. \tag{2.6}$$

For concreteness, we list some basic but practically important examples describing the true target and the proxy.

Example 2.8 (Integrated Variance): Let X_t be the continuous-time logarithmic price process of an asset, which is assumed to be an Itô semimartingale with the form $X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s + J_t$, where b_s is the stochastic drift, σ_s is the stochastic volatility, W is a Brownian motion and J is a pure-jump process. The integrated variance on day t is given by $IV_t = \int_{t-1}^t \sigma_s^2 ds$. If intraday observations on X_t are observable at sampling interval Δ, a popular proxy for IV_t is the realized variance estimator $RV_t = \sum_{i=1}^{[1/\Delta]} (\Delta_{t,i}X)^2$, where for each t and i, we denote $\Delta_{t,i}X = X_{(t-1)+i\Delta} - X_{(t-1)+(i-1)\Delta}$; see Andersen and Bollerslev (1998) and Andersen, Bollerslev, Diebold, and Labys (2003). In general, one can use jump-robust estimators such as the bipower variation $BV_t = \frac{\pi[1/\Delta]}{2([1/\Delta]-1)} \sum_{i=1}^{[1/\Delta]-1} |\Delta_{t,i}X||\Delta_{t,i+1}X|$ (Barndorff-Nielsen and Shephard (2004b)) or the truncated realized variance estimator $TV_t = \sum_{i=1}^{[1/\Delta]} (\Delta_{t,i}X)^2 \mathbf{1}\{|\Delta_{t,i}X| \le \bar{\alpha}\Delta^{\varpi}\}$ (Mancini

[9] A key motivation of our analysis is that while a high-frequency estimator of the latent variable is used by the forecaster for evaluation (and potentially estimation), the estimator is *not* the variable of interest. If the estimator is taken as the target variable, then no issues about the latency of the target variable arise, and existing predictive ability tests may be applied without modification. It is only in cases where the variable of interest is unobservable that further work is required to justify the use of an estimator of the latent target variable in predictive ability tests.

(2001)) as proxies for IV_t, where $\bar{\alpha} > 0$ and $\varpi \in (0, 1/2)$ are tuning parameters that specify the truncation threshold. In this case, the target to be forecast is $Y^{\dagger}_{t+\tau} = IV_{t+\tau}$ and the proxy $Y_{t+\tau}$ may be $RV_{t+\tau}$, $BV_{t+\tau}$ or $TV_{t+\tau}$.

Example 2.9 (Beta): Consider the same setting as in Example 2.8. Let M_t be the logarithmic price process of the market portfolio, which is also assumed to be an Itô semimartingale. In applications on hedging and portfolio management, it is of great interest to forecast the beta of the price process X_t with respect to the market portfolio. In a general continuous-time setting with price jumps, the beta of X_t can be defined as $[X, M]_t / [M, M]_t$, where $[\cdot, \cdot]_t$ denotes the quadratic covariation of two semimartingales over the time interval $[t-1, t]$; see Barndorff-Nielsen and Shephard (2004a). Here, $Y^{\dagger}_t$ is the beta on day t, which can be estimated by its realized counterpart $Y_t = \widehat{[X, M]}_t / \widehat{[M, M]}_t$, where $\widehat{[X, M]}_t = \sum_{i=0}^{[1/\Delta]} (\Delta_{t,i} X)(\Delta_{t,i} M)$ and $\widehat{[M, M]}_t$ is the realized variance of M over day t.

Our goal is to provide conditions under which the test ϕ_T introduced in Section 2.1 has the same asymptotic level and power properties under the true hypotheses, EPA and SPA, as it does under the proxy hypotheses, PEPA and PSPA. We achieve this by invoking Assumption C1, below, which we call an *approximation-of-hypothesis* condition.

Assumption C1: $a_T(\mathbb{E}[f^*_T] - \mathbb{E}[f^{\dagger}_T]) \to 0$, where a_T is given in Assumption A1.

Assumption C1 is clearly high-level. We provide more primitive conditions later in this subsection and devote Section 3 to providing concrete examples involving various estimators formed using high-frequency data. This presentation allows us to separate the main intuition behind the negligibility result, which is formalized by Theorem 2.1 below, from the somewhat technical calculations for high-frequency data.

Theorem 2.1: The statements of Proposition 2.1 hold with PEPA (resp. PSPA) replaced by EPA (resp. SPA), provided that Assumption C1 holds in addition.

Comments. (i) The negligibility result is achieved through the approximation of hypotheses, instead of the approximation of statistics. The latter approach may be carried out by showing that the approximation errors between $\bar{f}_T$, S_T, $z_{T,1-\alpha}$ and their "true" counterparts, i.e. statistics defined in the same way but with the proxy replaced by the true target variable, to be asymptotically negligible. An approximation-of-statistics approach would demand more structure on the auxiliary estimator S_T and the critical value estimator $z_{T,1-\alpha}$. As illustrated in the examples in Section 2.1, S_T and $z_{T,1-\alpha}$ may be constructed in very distinct ways even across the baseline applications considered there. The approximation-of-hypothesis argument conveniently allows one to be agnostic about the proxy error in S_T and $z_{T,1-\alpha}$, and hence agnostic about their structures. As a result, the negligibility result can be applied to the many apparently distinct settings described in Section 2.1, as well as some extensions described in Section 4.

(ii) The result established in Theorem 2.1 is a form of *weak* negligibility, in the sense that it only concerns the rejection probability. An alternative notion of negligibility can be framed as follows. Let $\phi_T^{\dagger}$ be a nonrandomized test that is constructed in the same way as ϕ_T but with $Y_{t+\tau}$ replaced by $Y_{t+\tau}^{\dagger}$. That is, $\phi_T^{\dagger}$ is the infeasible test we would use if we could observe the true forecast target. We may consider the difference between the proxy and the target negligible in a *strong* sense if $\mathbb{P}(\phi_T = \phi_T^{\dagger}) \to 1$. It is obvious that strong negligibility implies weak negligibility. While the strong negligibility may seem to be a reasonable result to pursue, we argue that the weak negligibility better suits, and is sufficient for, the testing context considered here. Strong negligibility requires the feasible and infeasible test decisions to agree, which may be too much to ask: for example, this would demand ϕ_T to equal $\phi_T^{\dagger}$ even if $\phi_T^{\dagger}$ commits a false rejection. Moreover, the strong negligibility would inevitably demand more assumptions and/or technical maneuvers, as noted in comment (i) above. Hence we do not pursue strong negligibility.

(iii) Similar to our negligibility result, West (1996) defines cases exhibiting "asymptotic irrelevance" as those in which valid inference about predictive ability can be made while ignoring the presence of parameter estimation error. Technically speaking, our negligibility result is very distinct from West's result: here, the unobservable quantity is a latent stochastic process $(Y_t^{\dagger})_{t\geq 1}$ that grows in T as $T \to \infty$, while in West's setting it is a fixed deterministic and finite-dimensional parameter β^*. That is, our asymptotic negligibility concerns a measurement error problem, while West's asymptotic irrelavance concerns, roughly speaking, a two-step estimation problem. Unlike West's (1996) case, where a correction can be applied when the asymptotic irrelevance condition (w.r.t. β^*) is not satisfied, no such correction (w.r.t. $Y_t^{\dagger}$) is readily available in our application. This is mainly because, in the setting of high-frequency financial econometrics with long-span data, an important component in the approximation error $Y_{t+\tau} - Y_{t+\tau}^{\dagger}$ is a *bias* term arising from the use of discretely sampled data for approximating the latent target that is defined in continuous time. Our approach shares the same nature as that of Corradi and Distaso (2006) and Todorov and Tauchen (2012b), although our econometric interest and content are very different from theirs; see Section 3.6 for further discussions.

We now consider sufficient conditions for Assumption C1. Below, Assumption C2 requires the proxy to be "precise." This assumption is still high-level and is further discussed in Section 3. Assumption C3 is a regularity-type condition. Detailed comments on these sufficient conditions are given below.

Assumption C2: There exist some bounded deterministic sequence $(d_t)_{t\geq 1}$ and constants $p \in [1, 2)$, $\theta > 0$, $C > 0$, such that $\|Y_{t+\tau} - Y_{t+\tau}^{\dagger}\|_p \leq C d_{t+\tau}^{\theta}$ for all $t \geq R$.

Assumption C3: (a) $\|f(Y_{t+\tau}, F_{t+\tau}(\beta^*)) - f(Y_{t+\tau}^{\dagger}, F_{t+\tau}(\beta^*))\| \leq m_{t+\tau}\|Y_{t+\tau} - Y_{t+\tau}^{\dagger}\|$ for some sequence $m_{t+\tau}$ of random variables and all $t \geq R$. Moreover, $\sup_{t \geq R} \|m_{t+\tau}\|_q < \infty$ for some $q \geq p/(p-1)$, where p is the same as in Assumption C2 and, for $p=1$, this condition is understood as $\sup_{t \geq R} \|m_{t+\tau}\| \leq C$ almost surely for some constant C.

(b) $(a_T/P) \sum_{t=R}^{T} d_{t+\tau}^{\theta} \to 0$, where a_T and θ are given in Assumptions A1 and C2, respectively.

LEMMA 2.1: Assumptions C2 and C3 imply Assumption C1.

COMMENTS ON ASSUMPTION C2. (i) In financial econometrics applications, d_t in Assumption C2 plays the role of a bound for the intraday sampling interval on day t. While more technical details are provided in Section 3, here we note that we allow d_t to vary across days. This flexibility is especially appealing when the empirical analysis involves a relatively long history of intraday data, because intraday data are typically sampled at lower frequencies in earlier periods than recent ones. The time-varying sampling scheme poses a challenge to existing inference methods on predictive accuracy, because these methods are often built under covariance stationarity (see, e.g. Diebold and Mariano (1995), West (1996) and White (2000)), although such restrictions are unlikely to be essential. The framework of Giacomini and White (2006) allows for data heterogeneity and naturally fits in our setting here. Giacomini and Rossi (2009) extend the theory of West (1996) to a heterogeneous setting, which is useful here.

(ii) Assumption C2 imposes an L_p bound on the approximation error of Y_t as a (typically fractional) polynomial of d_t. In many basic examples, this assumption holds for $\theta = 1/2$. See Section 3 for more details.

(iii) Assumption C2 is stable under linear transformations. A simple but practically important example is the subsampling-and-averaging method considered by Zhang, Mykland, and Aït-Sahalia (2005).[10] If $Y_{t+\tau}^{\dagger}$ has n_p proxies, say $(Y_{t+\tau}^{(j)})_{1 \leq j \leq n_p}$, and each of them satisfies Assumption C2, then their average $n_p^{-1} \sum_{j=1}^{n_p} Y_{t+\tau}^{(j)}$ also satisfies this assumption. Hence, the empirical worker can always "upgrade" a proxy using sparsely sampled data to its subsampled-and-averaged version so as to take advantage of all high-frequency data available while still being robust against market microstructure effects.

(iv) Assumption C2 is also preserved under certain nonlinear transformations, provided that additional moment conditions are properly imposed. For example, many economically interesting forecast targets, such as beta and correlation, are defined as ratios. To fix ideas, suppose that $Y_t^{\dagger} = A_t^{\dagger}/B_t^{\dagger}$. We consider a proxy $Y_t = A_t/B_t$ for $Y_t^{\dagger}$, where A_t and B_t are available proxies for $A_t^{\dagger}$ and $B_t^{\dagger}$ that verify $\|A_t - A_t^{\dagger}\|_{p'} + \|B_t - B_t^{\dagger}\|_{p'} \leq K d_t^{\theta}$ for some $p' \geq 1$. Let $p \in [1, p']$ and p'' satisfy $1/p' + 1/p'' = 1/p$. By the triangle inequality and Hölder's inequality, it is easy to see that $\|Y_t - Y_t^{\dagger}\|_p \leq \|1/B_t^{\dagger}\|_{p''} \|A_t - A_t^{\dagger}\|_{p'} + \|Y_t/B_t^{\dagger}\|_{p''} \|B_t - B_t^{\dagger}\|_{p'}$. Therefore, $\|Y_t - Y_t^{\dagger}\|_p \leq K d_t^{\theta}$

[10] In Zhang, Mykland, and Aït-Sahalia (2005), the estimand of interest is the integrated variance, but the scope of the idea of subsampling-and-averaging extends beyond integrated variance.

provided that $\|1/B_t^\dagger\|_{p''}$ and $\|Y_t/B_t^\dagger\|_{p''}$ are bounded; in particular, Y_t verifies Assumption C2. This calculation shows the benefit of considering a general L_p bound in Assumption C2.

Comments on Assumption C3. (i) Assumption C3(a) imposes smoothness of the evaluation function in the target variable. It is easily verified if $f(\cdot)$ collects pairwise loss differentials of competing forecasts. For example, if $f(Y,(F_1,F_2)) = L(Y-F_1) - L(Y-F_2)$ for some globally Lipschitz loss function $L(\cdot)$, then $m_{t+\tau}$ can be taken as a constant, and Assumption C3(a) holds trivially for $p=1$ (and, hence, any $p \geq 1$). An important example of such loss functions in the scalar setting is the lin-lin loss, i.e. $L(u) = (\gamma-1)\,u\mathbf{1}\{u<0\} + \gamma u\mathbf{1}\{u \geq 0\}$ for some asymmetry parameter $\gamma \in (0,1)$; this is the absolute error loss when $\gamma = 0.5$. Non-Lipschitz loss functions are also allowed. For example, when $L(u) = u^2$ (quadratic loss), Assumption C3(a) holds for $m_{t+\tau} = 2\,|F_{1,t+\tau}(\beta^*) - F_{2,t+\tau}(\beta^*)|$, provided that the forecasts have bounded moments up to order $p/(p-1)$; sometimes the forecasts are bounded by construction (e.g., forecasts of correlation coefficients), so we can again take $m_{t+\tau}$ to be a constant, and verify Assumption C3(a) for any $p \geq 1$.

(ii) Assumption C3(b) is a regularity condition that requires d_t to be sufficiently small in an average sense over the prediction sample. This condition formalizes the notion that a large sample not only includes more days, but also includes increasingly more intradaily observations. The asymptotic setting may be referred to as an “eventually fill-in” one. While the asymptotic embedding is not meant to be interpreted literally, it is interesting to note that this setting does mimic datasets seen in practice. This condition is relatively less restrictive when θ is large, that is, when a more accurate proxy is available, and vice versa.

(iii) The index p governs the trade-off between Assumptions C2 and C3. Assumption C2 (resp. C3) is stronger (resp. weaker) when p is higher, and vice versa. In particular, if $m_{t+\tau}$ in Assumption C3 can be taken bounded, then it is enough to verify Assumption C2 for $p=1$, which sometimes leads to better rates of convergence (i.e., higher values of θ). The main purpose of allowing $p>1$ is to allow some flexibility for verifying Assumption C3. For example, when $p=3/2$, we only need the L_q-boundedness condition in Assumption C3 to hold for $q=3$. Moment conditions of this sort are not strong, and often needed for other purposes in the theory of forecast evaluation, such as deriving a CLT or proving the consistency of a HAC estimator; see, e.g., Davidson (1994) and Andrews (1991).

3 Examples and primitive conditions for Assumption C2

This section provides several examples that verify the high-level assumption $\|Y_t - Y_t^\dagger\|_p \leq K d_t^\theta$ in Assumption C2 for some generic constant $K>0$. We consider a comprehensive list of latent risk measures defined as functionals of continuous-time volatility and jump processes, together with proxies formed using high-frequency data. Section 3.1 presents the setting and Sections 3.2–3.5 show the examples. In Section 3.6, we discuss the key technical differences between results here and those in the existing high-frequency literature.

3.1 Setup

We impose the following condition on the logarithmic price process, X_t:

Assumption HF: Fix a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. Suppose that the following conditions hold for constants $k \geq 2$ and $C > 0$.

(a) The process $(X_t)_{t\geq 0}$ is a d-dimensional Itô semimartingale with the following form

$$\begin{aligned} X_t &= X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s + J_t, \quad \text{where} \\ J_t &= \int_0^t \int_{\mathbb{R}} \delta(s,z) 1_{\{\|\delta(s,z)\|\leq 1\}} \tilde{\mu}(ds,dz) \\ &\quad + \int_0^t \int_{\mathbb{R}} \delta(s,z) 1_{\{\|\delta(s,z)\|> 1\}} \mu(ds,dz), \end{aligned} \tag{3.1}$$

and b is a d-dimensional càdlàg adapted process, W is a d'-dimensional standard Brownian motion, σ is a $d\times d'$ càdlàg adapted process, δ is a d-dimensional predictable function defined on $\Omega\times\mathbb{R}_+\times\mathbb{R}$, μ is a Poisson random measure on $\mathbb{R}_+ \times \mathbb{R}$ with compensator $\nu(ds,dz) = ds \otimes \lambda(dz)$ for some σ-finite measure λ, and $\tilde{\mu} = \mu - \nu$. We set $c_t = \sigma_t\sigma_t^\intercal$, that is, the spot covariance matrix.

(b) The process σ_t is a $d \times d'$ Itô semimartingale with the form

$$\sigma_t = \sigma_0 + \int_0^t \tilde{b}_s ds + \int_0^t \tilde{\sigma}_s dW_s + \int_0^t \int_{\mathbb{R}} \tilde{\delta}(s,z) \tilde{\mu}(ds,dz), \tag{3.2}$$

where $\tilde{b}$ is a $d \times d'$ càdlàg adapted process, $\tilde{\sigma}$ is a $d \times d' \times d'$ càdlàg adapted process and $\tilde{\delta}(\cdot)$ is a $d \times d'$ predictable function on $\Omega \times \mathbb{R}_+ \times \mathbb{R}$.

(c) For some constant $r \in (0,2]$, and nonnegative deterministic functions $\Gamma(\cdot)$ and $\tilde{\Gamma}(\cdot)$ on $\mathbb{R}$, we have $\|\delta(\omega,s,z)\| \leq \Gamma(z)$ and $\|\tilde{\delta}(\omega,s,z)\| \leq \tilde{\Gamma}(z)$ for all $(\omega,s,z) \in \Omega \times \mathbb{R}_+ \times \mathbb{R}$ and

$$\begin{aligned} &\int_{\mathbb{R}} (\Gamma(z)^r \wedge 1)\lambda(dz) + \int_{\mathbb{R}} \Gamma(z)^k 1_{\{\Gamma(z)>1\}}\lambda(dz) < \infty, \\ &\int_{\mathbb{R}} (\tilde{\Gamma}(z)^2 + \tilde{\Gamma}(z)^k)\lambda(dz) < \infty. \end{aligned} \tag{3.3}$$

(d) Let $b'_s = b_s - \int_{\mathbb{R}} \delta(s,z) 1_{\{\|\delta(s,z)\|\leq 1\}}\lambda(ds)$ if $r \in (0,1]$ and $b'_s = b_s$ if $r \in (1,2]$. We have for all $s \geq 0$,

$$\mathbb{E}\|b'_s\|^k + \mathbb{E}\|\sigma_s\|^k + \mathbb{E}\|\tilde{b}_s\|^k + \mathbb{E}\|\tilde{\sigma}_s\|^k \leq C. \tag{3.4}$$

(e) For each day t, the process X is sampled at deterministic discrete times $t-1 = \tau(t,0) < \cdots < \tau(t, n_t) = t$, where n_t is the number of intraday returns. Moreover, with $d_{t,i} = \tau(t,i) - \tau(t,i-1)$, we have $d_t = \sup_{1\leq i\leq n_t} d_{t,i} \to 0$ and $n_t = O(d_t^{-1})$ as $t \to \infty$.

Parts (a) and (b) in Assumption HF are standard in the study of high-frequency data, which require the price X_t and the stochastic volatility process σ_t to be Itô semimartingales. Part (c) imposes a type of dominance condition on the random jump size for the price and the volatility. The constant r governs the concentration of small jumps, as it provides an upper bound for the generalized Blumenthal-Getoor index. The integrability condition in part (c) is weaker when r is larger. The kth-order integrability of $\Gamma(\cdot)\mathbf{1}\{\Gamma(\cdot) > 1\}$ and $\tilde{\Gamma}(\cdot)$ with respect to the intensity measure λ is needed to facilitate the derivation of bounds via sufficiently high moments; these are restrictions on "big" jumps. Part (d) imposes integrability conditions to serve the same purpose.[11] Part (e) describes the sampling scheme of the intraday data. As mentioned in comment (i) of Assumption C2, we allow X to be sampled at irregular times with the mesh d_t going to zero "eventually" in later samples.

Below, for each $t \geq 1$ and $i \geq 1$, we denote the ith return of X in day t by $\Delta_{t,i}X$, i.e. $\Delta_{t,i}X = X_{\tau(t,i)} - X_{\tau(t,i-1)}$.

3.2 Generalized realized variations for continuous processes

We start with the basic setting with X continuous. Consider the following general class of estimators: for any function $g : \mathbb{R}^d \mapsto \mathbb{R}$, we set $\widehat{\mathcal{I}}_t(g) \equiv \sum_{i=1}^{n_t} g(\Delta_{t,i}X/d_{t,i}^{1/2})d_{t,i}$; recall that $d_{t,i}$ is the length of the sampling interval associated with the return $\Delta_{t,i}X$. We also associate with g the following function: for any $d \times d$ positive semidefinite matrix A, we set $\rho(A; g) = \mathbb{E}[g(U)]$ for $U \sim \mathcal{N}(0, A)$, provided that the expectation is well-defined. Proposition 3.1 below provides a bound for the approximation error of the proxy $\widehat{\mathcal{I}}_t(g)$ relative to the target variable $\mathcal{I}_t(g) \equiv \int_{t-1}^{t} \rho(c_s; g)\, ds$. In the notation from Section 2, $\widehat{\mathcal{I}}_t(g)$ corresponds to the proxy Y_t and $\mathcal{I}_t(g)$ to the latent target variable $Y_t^{\dagger}$.

PROPOSITION 3.1: Let $p \in [1, 2)$. For some constant $C > 0$, suppose the following conditions hold: (i) X_t is continuous; (ii) $g(\cdot)$ and $\rho(\cdot\,; g)$ are continuously differentiable and, for some $q \geq 0$, $\|\partial_x g(x)\| \leq C(1 + \|x\|^q)$ and $\|\partial_A \rho(A; g)\| \leq C(1 + \|A\|^{q/2})$; (iii) Assumption HF with $k \geq \max\{2qp/(2-p), 4\}$; (iv) $\mathbb{E}[\rho(c_s; g^2)] \leq C$ for all $s \geq 0$. Then $\|\widehat{\mathcal{I}}_t(g) - \mathcal{I}_t(g)\|_p \leq K d_t^{1/2}$.

[11] The kth-order integrability conditions in Assumptions HF(c,d) are imposed explicitly because we are interested in an asymptotic setting with the time span $T \to \infty$, which is very different from the fill-in asymptotic setting with fixed time span. In the latter case, one can invoke the classical localization argument and assume that Γ, $\tilde{\Gamma}$, b_s, b'_s, σ_s, $\tilde{\sigma}_s$ and $\tilde{b}_s$ to be uniformly bounded without loss of generality when proving limit theorems and deriving stochastic bounds; the uniform boundedness then trivially implies the integrability conditions in parts (c) and (d) in Assumption HF.

COMMENT. In many applications, the function $\rho\left(\,\cdot\,; g\right)$ can be expressed in closed form. For example, if we take $g\left(x\right) = \left|x\right|^{a}/m_a$ for some $a \geq 2$, where $x \in \mathbb{R}$ and m_a is the ath absolute moment of a standard Gaussian variable, then $\mathcal{I}_t(g) = \int_{t-1}^{t} c_s^{a/2} ds$. Another univariate example is to take $g(x) = \cos(\sqrt{2u}x)$, yielding $\mathcal{I}_t(g) = \int_{t-1}^{t} \exp(-uc_s)ds$. In this case, $\widehat{\mathcal{I}}_t(g)$ is the realized Laplace transform of volatility (Todorov and Tauchen (2012b)) and $\mathcal{I}_t(g)$ is the Laplace transform of the volatility occupation density (Todorov and Tauchen (2012a), Li, Todorov, and Tauchen (2012)). A simple bivariate example is $g(x_1, x_2) = x_1 x_2$, which leads to $\mathcal{I}_t(g) = \int_{t-1}^{t} c_{12,s} ds$, that is, the integrated covariance between the two components of X_t.

3.3 Functionals of price jumps

In this subsection, we consider target variables that are functionals of the jumps of X. We denote $\Delta X_t = X_t - X_{t-}$, $t \geq 0$. The functional of interest has the form $\mathcal{J}_t(g) \equiv \sum_{t-1<s\leq t} g\left(\Delta X_s\right)$ for some function $g : \mathbb{R}^d \mapsto \mathbb{R}$. The proxy is the sample analogue estimator: $\widehat{\mathcal{J}}_t\left(g\right) \equiv \sum_{i=1}^{n_t} g\left(\Delta_{t,i} X\right)$.

PROPOSITION 3.2: Let $p \in [1, 2)$. Suppose (i) g is twice continuously differentiable; (ii) for some $q_2 \geq q_1 \geq 3$ and a constant $C > 0$, we have $\|\partial_x^j g(x)\| \leq C(\|x\|^{q_1-j} + \|x\|^{q_2-j})$ for all $x \in \mathbb{R}^d$ and $j \in \{0, 1, 2\}$; (iii) Assumption HF with $k \geq \max\{2q_2, 4p/(2-p)\}$. Then $\|\widehat{\mathcal{J}}_t\left(g\right) - \mathcal{J}_t(g)\|_p \leq K d_t^{1/2}$.

COMMENTS. (i) The polynomial $\|x\|^{q_1-j}$ in condition (ii) bounds the growth of $g(\cdot)$ and its derivatives near zero. This condition ensures that the contribution of the continuous part of X to the approximation error is dominated by the jump part of X. This condition can be relaxed at the cost of a more complicated expression for the rate. The polynomial $\|x\|^{q_2-j}$ controls the growth of $g(\cdot)$ near infinity so as to tame the effect of big jumps.

(ii) Basic examples include unnormalized realized skewness ($g(x) = x^3$), kurtosis ($g(x) = x^4$), coskewness ($g(x_1, x_2) = x_1^2 x_2$) and cokurtosis ($g(x_1, x_2) = x_1^2 x_2^2$).[12] Bounds on the proxy accuracy of their normalized counterparts can then be obtained following comment (iv) of Assumption C2. See Amaya, Christoffersen, Jacobs, and Vasquez (2011) for applications using these risk measures.

3.4 Jump-robust volatility functionals

In this subsection, we consider a general class of volatility functionals with proxies that are robust to jumps in X. Let $g : \mathbb{R}^{d\times d} \mapsto \mathbb{R}$ be a continuous function. The volatility functional of interest is $\mathcal{I}_t^{\star}(g) = \int_{t-1}^{t} g(c_s)ds$. So as to construct the jump-robust proxy for $\mathcal{I}_t^{\star}(g)$, we first nonparametrically recover the spot covariance process by using a local truncated variation estimator

$$\hat{c}_{\tau(t,i)} = \frac{1}{k_t} \sum_{j=1}^{k_t} d_{t,i+j}^{-1} \Delta_{t,i+j} X \Delta_{t,i+j} X^{\intercal} 1_{\{\|\Delta_{t,i+j} X\| \leq \bar{\alpha} d_{t,i+j}^{\varpi}\}}, \tag{3.5}$$

[12] Under the fill-in asymptotic setting with fixed span, the unnormalized (co)skewness does not admit a central limit theorem; see the comment following Theorem 5.1.2 in Jacod and Protter (2012), p. 128.

where $\bar{\alpha} > 0$ and $\varpi \in (0, 1/2)$ are constant tuning parameters, and k_t denotes a sequence of integers that specifies the local window for the spot covariance estimation. The proxy for $\mathcal{I}_t^\star(g)$ is the sample analogue estimator $\widehat{\mathcal{I}}_t^\star(g) = \sum_{i=0}^{n_t - k_t} g(\hat{c}_{\tau(t,i)}) d_{t,i}$.

PROPOSITION 3.3: Let $q \geq 2$ and $p \in [1, 2)$ be constant. Suppose (i) g is twice continuously differentiable and $\|\partial_x^j g(x)\| \leq C(1+\|x\|^{q-j})$ for $j = 0, 1, 2$ and some constant $C > 0$; (ii) $k_t \asymp d_t^{-1/2}$; (iii) Assumption HF with $k \geq \max\{4q, 4p(q-1)/(2-p), (1-\varpi r)/(1/2-\varpi)\}$ and $r \in (0, 2)$. We set $\theta_1 = 1/(2p)$ in the general case and $\theta_1 = 1/2$ if we further assume σ_t is continuous. We also set $\theta_2 = \min\{1 - \varpi r + q(2\varpi - 1), 1/r - 1/2\}$. Then $\|\widehat{\mathcal{I}}_t^\star(g) - \mathcal{I}_t^\star(g)\|_p \leq K d_t^{\theta_1 \wedge \theta_2}$.

COMMENTS. (i) The rate exponent θ_1 is associated with the contribution from the continuous component of X_t. The exponent θ_2 captures the approximation error due to the elimination of jumps. If we further impose $r < 1$ and $\varpi \in [(q - 1/2)/(2q - r), 1/2)$, then $\theta_2 \geq 1/2 \geq \theta_1$. That is, the presence of "inactive" jumps does not affect the rate of convergence, provided that the jumps are properly truncated.

(ii) Jacod and Rosenbaum (2012) characterize the limit distribution of $\widehat{\mathcal{I}}_t^\star(g)$ under the fill-in asymptotic setting with fixed span, under the assumption that g is three-times continuously differentiable and $r < 1$. Here, we obtain the same rate of convergence under the L_1 norm, and under the L_p norm if σ_t is continuous, in the eventually fill-in setting with $T \to \infty$. Our results also cover the case with active jumps, that is, the setting with $r \geq 1$.

3.5 Additional special examples

We now consider a few special examples which are not covered by Propositions 3.1–3.3. In the first example, the true target is the daily quadratic variation matrix QV_t of the process X, that is, $QV_t = \int_{t-1}^t c_s ds + \sum_{t-1<s\leq t} \Delta X_s \Delta X_s^\intercal$. The associated proxy is the realized covariance matrix $RV_t \equiv \sum_{i=1}^{n_t} \Delta_{t,i} X \Delta_{t,i} X^\intercal$.

PROPOSITION 3.4: Let $p \in [1, 2)$. Suppose Assumption HF with $k \geq \max\{2p/(2-p), 4\}$. Then $\|RV_t - QV_t\|_p \leq K d_t^{1/2}$.

Second, we consider the bipower variation of Barndorff-Nielsen and Shephard (2004b) for univariate X that is defined as

$$BV_t = \frac{n_t}{n_t - 1} \frac{\pi}{2} \sum_{i=1}^{n_t - 1} |d_{t,i}^{-1/2} \Delta_{t,i} X| |d_{t,i+1}^{-1/2} \Delta_{t,i+1} X| d_{t,i}. \tag{3.6}$$

This estimator serves as a proxy for the integrated variance $\int_{t-1}^t c_s ds$.

PROPOSITION 3.5: Let $1 \leq p < p' \leq 2$. Suppose that Assumption HF holds with $d = 1$ and $k \geq \max\{pp'/(p'-p), 4\}$. We have (a) $\|BV_t - \int_{t-1}^t c_s ds\|_p \leq K d_t^{(1/r) \wedge (1/p') - 1/2}$; (b) if, in addition, X is continuous, then $\|BV_t - \int_{t-1}^t c_s ds\|_p \leq K d_t^{1/2}$.

COMMENT. Part (b) shows that, when X is continuous, the approximation error of the bipower variation achieves the $\sqrt{n_t}$ rate. Part (a) provides a bound for the rate of convergence (under L_p) in the case with jumps. The rate is slower than that in the continuous case. Not surprisingly, the rate is sharper if r is smaller (i.e., jumps are less active), and p and p' are close to 1. In particular, with $r \leq 1$ and p' being close to 1, the bound in the jump case can be made arbitrarily close to $O(d_t^{1/2})$, at the cost of assuming higher-order moments to be finite (i.e., larger k). The slower rate in the jump case is in line with the known fact that the bipower variation estimator does not admit a CLT when X is discontinuous.[13]

Finally, we consider the realized semivariance estimator proposed by Barndorff-Nielsen, Kinnebrouck, and Shephard (2010) for univariate X. Let $\{x\}_+$ and $\{x\}_-$ denote the positive and the negative parts of $x \in \mathbb{R}$, respectively. The upside $(+)$ and the downside $(-)$ realized semivariances are defined as $\widehat{SV}_t^{\pm} = \sum_{i=1}^{n_t} \{\Delta_{t,i} X\}_{\pm}^2$, which serve as proxies for $SV_t^{\pm} = \frac{1}{2}\int_{t-1}^{t} c_s ds + \sum_{t-1<s\leq t} \{\Delta X_s\}_{\pm}^2$.

PROPOSITION 3.6: Let $1 \leq p < p' \leq 2$. Suppose that Assumption HF holds with $d = 1$, $r \in (0, 1]$ and $k \geq \max\{pp'/(p'-p), 4\}$. Then (a) $\|\widehat{SV}_t^{\pm} - SV_t^{\pm}\|_p \leq K d_t^{1/p'-1/2}$; (b) if, in addition, X is continuous, then $\|\widehat{SV}_t^{\pm} - SV_t^{\pm}\|_p \leq K d_t^{1/2}$.

COMMENT. Part (b) shows that, when X is continuous, the approximation error of the semivariance achieves the $\sqrt{n_t}$ rate, which agrees with the rate shown in Barndorff-Nielsen, Kinnebrouck, and Shephard (2010), but in a different asymptotic setting. Part (a) provides a bound for the rate of convergence in the case with jumps. The constant p' arises as a technical device in the proof. One should make it small so as to achieve a better rate, subject to the regularity condition $k \geq pp'/(p'-p)$. In particular, the rate can be made close to that in the continuous case when p', hence p too, are close to 1. Barndorff-Nielsen, Kinnebrouck, and Shephard (2010) do not consider rate results in the case with price or volatility jumps.

3.6 Technical comments

The proofs of Propositions 3.1–3.6 are based on standard techniques reviewed and developed in Jacod and Protter (2012). That being said, these results are distinct from those in Jacod and Protter (2012), and those in the original papers that propose the above estimators, in several aspects. First, existing results on the rate of convergence in the fill-in setting with fixed span cannot be directly invoked here because we consider a setting with $T \to \infty$. Technically speaking, the localization argument (see Section 4.4.1 in Jacod and Protter (2012)) cannot be invoked and the proofs demand extra care. Second, we are only interested in the rate of convergence, rather

[13] See p. 313 in Jacod and Protter (2012) and Vetter (2010).

than proving a central limit theorem. We thus provide direct proofs on the rates, which are (sometimes much) shorter than proofs of central limit theorems for the high-frequency estimators. Third, bounding the L_p norm of the approximation error is our pursuit here; see comment (iv) of Assumption C2 and comment (iii) of Assumption C3 for its usefulness. However, the L_p bound is typically not of direct interest in the proof of limit theorems, where one is mainly concerned with establishing convergence in probability (see Theorem IX.7.28 in Jacod and Shiryaev (2003)) and establishing stochastic orders. Finally, we note that for estimators with known central limit theorems under the fill-in asymptotic setting, we establish the $\sqrt{n_t}$ rate of convergence under the L_p norm. Moreover, we also provide rate results for estimators that do not have known central limit theorems; examples include the bipower variation and the semivariance when there are price jumps, realized (co)skewness, and jump-robust estimators for volatility functionals under the setting with active jumps, to name a few.

Papers that consider bounds for high-frequency proxy errors under the double asymptotic setting, that is, the setting with both the time span and the number of intraday observations going to infinity, include Corradi and Distaso (2006) and Todorov and Tauchen (2012b).

Corradi and Distaso (2006), followed by Corradi, Distaso, and Swanson (2009, 2011), consider proxies for the quadratic variation, including the realized variance, the bipower variation, and noise-robust estimators such as the multiscale realized variance (Zhang, Mykland, and Aït-Sahalia (2005), Zhang (2006)) and the realized kernel (Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008)). In the absence of microstructure noise, Propositions 3.1, 3.4 and 3.5 complement these existing results by considering the case with general price jumps without assuming jumps to have finite activity as considered by Corradi, Distaso, and Swanson (2011). This generalization is empirically relevant in view of the findings of Aït-Sahalia and Jacod (2009, 2012). We stress that our main technical contribution is to consider a comprehensive list of high-frequency proxies that is well beyond the basic case of quadratic variation. We do not consider microstructure noise in this paper. However, we do allow for the subsampled-and-averaged realized variance estimator of Zhang, Mykland, and Aït-Sahalia (2005). Indeed, we can apply the subsampling-and-averaging technique to any proxy, as noted in comment (iii) of Assumption C2.

Todorov and Tauchen (2012b) consider the approximation error of the realized Laplace transform of volatility as a proxy for the Laplace transform $\int_{t-1}^{t}\exp(-u\sigma_s^2)ds$, $u>0$, of the volatility occupation density. In the absence of price jumps, the realized Laplace transform is a special case of Proposition 3.1. Todorov and Tauchen (2012b) allow for finite-variational price jumps, which is not considered in Proposition 3.1. That being said, an alternative proxy of $\int_{t-1}^{t}\exp(-u\sigma_s^2)ds$ is given by $\widehat{\mathcal{I}}_t^{\star}(g)$ with $g(x)=\exp(-ux)$, and Proposition 3.3 provides an L_p bound for the proxy error under a setting with possibly infinite-variational jumps.

4 Extensions: additional forecast evaluation methods

In this section we discuss several extensions of our baseline negligibility result (Theorem 2.1). We first consider tests for instrumented conditional moment equalities, as in Giacomini and White (2006). We then consider stepwise evaluation procedures that include the multiple testing method of Romano and Wolf (2005) and the model confidence set of Hansen, Lunde, and Nason (2011). Our purpose is twofold: one is to facilitate the application of these methods in the context of forecasting latent risk measures, the other is to demonstrate the generalizability of the framework developed so far through known, but distinct, examples. The two stepwise procedures each involve some method-specific aspects that are not used elsewhere in the paper; hence, for the sake of readability, we only briefly discuss the results here, and present the details (assumptions, algorithms and formal results) in the Supplemental Material.

4.1 Tests for instrumented conditional moment equalities

Many interesting forecast evaluation problems can be stated as a test for the conditional moment equality:

$$H_0 : \mathbb{E}[g(Y^{\dagger}_{t+\tau}, F_{t+\tau}(\beta^*))|\mathcal{H}_t] = 0, \quad \text{all } t \geq 0, \tag{4.1}$$

where $\mathcal{H}_t$ is a sub-σ-field that represents the forecast evaluator's information set at day t, and $g(\cdot,\cdot) : \mathcal{Y} \times \mathcal{Y}^{\bar{k}} \mapsto \mathbb{R}^{\kappa_g}$ is a measurable function. Specific examples are given below. Let h_t denote a $\mathcal{H}_t$-measurable $\mathbb{R}^{\kappa_h}$-valued data sequence that serves as an instrument. The conditional moment equality (4.1) implies the following unconditional moment equality:

$$H_{0,h} : \mathbb{E}[g(Y^{\dagger}_{t+\tau}, F_{t+\tau}(\beta^*)) \otimes h_t] = 0, \quad \text{all } t \geq 0. \tag{4.2}$$

We cast (4.2) in the setting of Section 2 by setting $f(Y_{t+\tau}, F_{t+\tau}(\beta^*), h_t) = g(Y_{t+\tau}, F_{t+\tau}(\beta^*)) \otimes h_t$. Then the theory in Section 2 can be applied without further change. In particular, Theorem 2.1 suggests that the two-sided PEPA test (with $\chi = 0$) using the proxy has a valid asymptotic level under H_0 and is consistent against the alternative

$$H_{2a,h} : \liminf_{T\to\infty} \|\mathbb{E}[g(Y^{\dagger}_{t+\tau}, F_{t+\tau}(\beta^*)) \otimes h_t]\| > 0. \tag{4.3}$$

Examples include tests for conditional predictive accuracy and tests for conditional forecast rationality. To simplify the discussion, we only consider scalar forecasts, so $\kappa_Y = 1$. Below, let $L(\cdot,\cdot) : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ be a loss function, with its first and second arguments being the target and the forecast.

Example 4.1: Giacomini and White (2006) consider two-sided tests for conditional equal predictive ability of two sequences of actual forecasts $F_{t+\tau} = (F_{1,t+\tau}, F_{2,t+\tau})$. The null hypothesis of interest is (4.1) with $g(Y^{\dagger}_{t+\tau}, F_{t+\tau}(\beta^*)) = L(Y^{\dagger}_{t+\tau}, F_{1,t+\tau}(\beta^*)) - L(Y^{\dagger}_{t+\tau}, F_{2,t+\tau}(\beta^*))$.

Since Giacomini and White (2006) concern the actual forecasts, we set β^* to be empty and treat $F_{t+\tau} = (F_{1,t+\tau}, F_{2,t+\tau})$ as an observable sequence. Primitive conditions for Assumptions A1 and A3 can be found in Giacomini and White (2006), which involve standard regularity conditions for weak convergence and HAC estimation. The test statistic is of Wald-type and readily verifies Assumptions A2 and B2. As noted by Giacomini and White (2006), their test is consistent against the alternative (4.3) and the power generally depends on the choice of h_t.

EXAMPLE 4.2: The population forecast $F_{t+\tau}(\beta^*)$, which is also the actual forecast if β^* is empty, is rational with respect to the information set $\mathcal{H}_t$ if it solves $\min_{F\in\mathcal{H}_t} \mathbb{E}[L(Y^\dagger_{t+\tau}, F)|\mathcal{H}_t]$ almost surely. Suppose that $L(y, F)$ is differentiable in F for almost every $y \in \mathcal{Y}$ under the conditional law of $Y^\dagger_{t+\tau}$ given $\mathcal{H}_t$, with the partial derivative denoted by $\partial_F L(\cdot, \cdot)$. As shown in Patton and Timmermann (2010), a test for conditional rationality can be carried out by testing the first-order condition of the minimization problem. That is to test the null hypothesis (4.1) with $g(Y^\dagger_{t+\tau}, F_{t+\tau}(\beta^*)) = \partial_F L(Y^\dagger_{t+\tau}, F_{t+\tau}(\beta^*))$. The variable $g(Y^\dagger_{t+\tau}, F_{t+\tau}(\beta^*))$ is the generalized forecast error (Granger (1999)). In particular, when $L(y, F) = (F - y)^2/2$, that is, the quadratic loss, we have $g(Y^\dagger_{t+\tau}, F_{t+\tau}(\beta^*)) = F - y$; in this case, the test for conditional rationality is reduced to a test for conditional unbiasedness. Tests for unconditional rationality and unbiasedness are special cases of their conditional counterparts, with $\mathcal{H}_t$ being the degenerate information set.

4.2 Stepwise multiple testing procedure for superior predictive accuracy

In the context of forecast evaluation, the multiple testing problem of Romano and Wolf (2005) consists of $\bar{k}$ individual testing problems of pairwise comparison for superior predictive accuracy. Let $F_{0,t+\tau}(\cdot)$ be the benchmark forecast model and let $f^\dagger_{j,t+\tau} = L(Y^\dagger_{t+\tau}, F_{0,t+\tau}(\beta^*)) - L(Y^\dagger_{t+\tau}, F_{j,t+\tau}(\beta^*))$, $1 \le j \le \bar{k}$, be the relative performance of forecast j relative to the benchmark. As before, $f^\dagger_{j,t+\tau}$ is defined using the true target variable $Y^\dagger_{t+\tau}$. We consider $\bar{k}$ pairs of hypotheses

$$\text{Multiple SPA} \begin{cases} H_{j,0} : \mathbb{E}[f^\dagger_{j,t+\tau}] \le 0 \text{ for all } t \ge 1, \\ H_{j,a} : \liminf_{T\to\infty} \mathbb{E}[\bar{f}^\dagger_{j,T}] > 0, \end{cases} \qquad 1 \le j \le \bar{k}. \tag{4.4}$$

These hypotheses concern the true target variable and are stated to allow for data heterogeneity.

Romano and Wolf (2005) propose a stepwise multiple (StepM) testing procedure that conducts decisions for individual testing problems while asymptotically control the familywise error rate (FWE), that is, the probability of any null hypothesis being falsely rejected. The StepM procedure relies on the observability of the forecast target. By imposing the approximation-of-hypothesis condition (Assumption C1), we can show that the StepM procedure, when applied to the proxy, asymptotically controls the FWE for the hypotheses (4.4) that concern the latent target. The details are in Supplemental Appendix S.B.1.

4.3 Model confidence sets

The *model confidence set* (MCS) proposed by Hansen, Lunde, and Nason (2011), henceforth HLN, can be specialized in the forecast evaluation context to construct confidence sets for superior forecasts. To fix ideas, let $f^{\dagger}_{j,t+\tau}$ denote the performance (e.g., the negative loss) of forecast j with respect to the true target variable. The set of superior forecasts is defined as

$$\overline{\mathcal{M}}^{\dagger} \equiv \left\{ j \in \{1, \ldots, \bar{k}\} : \mathbb{E}[f^{\dagger}_{j,t+\tau}] \geq \mathbb{E}[f^{\dagger}_{l,t+\tau}] \text{ for all } 1 \leq l \leq \bar{k} \text{ and } t \geq 1 \right\}.$$

That is, $\overline{\mathcal{M}}^{\dagger}$ collects the forecasts that are superior to others when evaluated using the true target variable. Similarly, the set of asymptotically inferior forecasts is defined as

$$\begin{aligned}\underline{\mathcal{M}}^{\dagger} \equiv \Big\{ j \in \{1, \ldots, \bar{k}\} : \ & \liminf_{T\to\infty} \left(\mathbb{E}[f^{\dagger}_{l,t+\tau}] - \mathbb{E}[f^{\dagger}_{j,t+\tau}] \right) > 0 \\ & \text{for some (and hence any) } l \in \overline{\mathcal{M}}^{\dagger} \Big\}.\end{aligned}$$

We are interested in constructing a sequence $\widehat{\mathcal{M}}_{T,1-\alpha}$ of $1-\alpha$ nominal level MCS's for $\overline{\mathcal{M}}^{\dagger}$ so that

$$\liminf_{T\to\infty} \left(\overline{\mathcal{M}}^{\dagger} \subseteq \widehat{\mathcal{M}}_{T,1-\alpha} \right) \geq 1 - \alpha, \quad \mathbb{P}\left(\widehat{\mathcal{M}}_{T,1-\alpha} \cap \underline{\mathcal{M}}^{\dagger} = \emptyset \right) \to 1. \tag{4.5}$$

That is, $\widehat{\mathcal{M}}_{T,1-\alpha}$ has valid (pointwise) asymptotic coverage and has asymptotic power one against fixed alternatives.

HLN's theory for the MCS is not directly applicable due to the latency of the forecast target. Following the prevailing strategy of the current paper, we propose a feasible version of HLN's algorithm that uses the proxy in place of the associated latent target. Under Assumption C1, we can show that this feasible version achieves (4.5). The details are in Supplemental Appendix S.B.2.

5 Monte Carlo analysis

5.1 Simulation designs

We consider three simulation designs which are intended to cover some of the most common and important applications of high-frequency data in forecasting: (A) forecasting univariate volatility in the absence of price jumps; (B) forecasting univariate volatility in the presence of price jumps; and (C) forecasting correlation. In each design, we consider the EPA hypotheses (2.5) under the quadratic loss for two competing one-day-ahead forecasts using the method of Giacomini and White (2006).

Each forecast is formed using a rolling scheme with window size $R \in \{500, 1000\}$ days. The prediction sample contains $P \in \{500, 1000, 2000\}$ days. The high-frequency data are simulated using the Euler scheme at every second, and proxies are computed using sampling interval $\Delta = 5$

seconds, 1 minute, 5 minutes, or 30 minutes. Each day contains 6.5 trading hours. There are 250 Monte Carlo trials in each experiment. All tests are at the 5% nominal level.

We now describe the simulation designs. Simulation A concerns forecasting the logarithm of the quadratic variation of a continuous price process. Following one of the simulation designs in Andersen, Bollerslev, and Meddahi (2005), we simulate the logarithmic price X_t and the spot variance process σ_t^2 according to the following stochastic differential equations:

$$\begin{cases} dX_t = 0.0314dt + \sigma_t(-0.576dW_{1,t} + \sqrt{1-0.576^2}dW_{2,t}) + dJ_t, \\ d\log\sigma_t^2 = -0.0136(0.8382 + \log\sigma_t^2)dt + 0.1148dW_{1,t}, \end{cases} \tag{5.1}$$

where W_1 and W_2 are independent Brownian motions and the jump process J is set to be identically zero. The target variable to be forecast is $\log IV_t$ and the proxy is $\log RV_t^{\Delta}$ with $\Delta = 5$ seconds, 1 minute, 5 minutes, or 30 minutes; recall Example 2.8 for the definitions of IV_t and RV_t^{Δ}.

The first forecast model in Simulation A is a GARCH(1,1) model (Bollerslev (1986)) estimated using quasi maximum likelihood on daily returns:

$$\text{Model A1:} \quad \begin{cases} r_t = X_t - X_{t-1} = \sigma_t\varepsilon_t, \quad \varepsilon_t|\mathcal{F}_{t-1} \sim \mathcal{N}(0,1), \\ \sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha r_{t-1}^2. \end{cases} \tag{5.2}$$

The second model is a heterogeneous autoregressive (HAR) model (Corsi (2009)) for $RV_t^{5\text{min}}$ estimated via ordinary least squares:

$$\text{Model A2:} \quad \begin{cases} RV_t^{5\text{min}} = \beta_0 + \beta_1 RV_{t-1}^{5\text{min}} + \beta_2\sum_{k=1}^{5} RV_{t-k}^{5\text{min}} \\ \qquad\qquad +\beta_3\sum_{k=1}^{22} RV_{t-k}^{5\text{min}} + e_t. \end{cases} \tag{5.3}$$

The logarithm of the one-day-ahead forecast for σ_{t+1}^2 (resp. $RV_{t+1}^{5\text{min}}$) from the GARCH (resp. HAR) model is taken as a forecast for $\log IV_{t+1}$.

In Simulation B, we also set the forecast target to be $\log IV_t$, but consider a more complicated setting with price jumps. We simulate X_t and σ_t^2 according to (5.1) and, following Huang and Tauchen (2005), we specify J_t as a compound Poisson process with intensity $\lambda = 0.05$ per day and with jump size distribution $\mathcal{N}(0.2, 1.4^2)$. The proxy for IV_t is the bipower variation BV_t^{Δ}; recall Example 2.8 for definitions.

The competing forecast sequences in Simulation B are as follows. The first forecast is based on a simple random walk model, applied to the 5-minute bipower variation $BV_t^{5\text{min}}$:

$$\text{Model B1:} \quad BV_t^{5\text{min}} = BV_{t-1}^{5\text{min}} + \varepsilon_t, \quad \text{where} \quad \mathbb{E}[\varepsilon_t|\mathcal{F}_{t-1}] = 0. \tag{5.4}$$

The second model is a HAR model for $BV_t^{1\text{min}}$

$$\text{Model B2:} \quad \begin{cases} BV_t^{1\text{min}} = \beta_0 + \beta_1 BV_{t-1}^{1\text{min}} + \beta_2\sum_{k=1}^{5} BV_{t-k}^{1\text{min}} \\ \qquad\qquad +\beta_3\sum_{k=1}^{22} BV_{t-k}^{1\text{min}} + e_t. \end{cases} \tag{5.5}$$

The logarithm of the one-day-ahead forecast for $BV_{t+1}^{5\,\text{min}}$ (resp. $BV_{t+1}^{1\,\text{min}}$) from the random walk (resp. HAR) model is taken as a forecast for $\log IV_{t+1}$.

Finally, we consider correlation forecasting in Simulation C. This simulation exercise is of particular interest as our empirical application in Section 6 concerns a similar forecasting problem. We adopt the bivariate stochastic volatility model used in the simulation study of Barndorff-Nielsen and Shephard (2004a). Let $W_t = (W_{1,t}, W_{2,t})$. The bivariate logarithmic price process X_t is given by

$$dX_t = \sigma_t dW_t, \quad \sigma_t \sigma_t^\intercal = \begin{pmatrix} \sigma_{1,t}^2 & \rho_t \sigma_{1,t}\sigma_{2,t} \\ \bullet & \sigma_{2,t}^2 \end{pmatrix}.$$

Let $B_{j,t}$, $j = 1, \ldots, 4$, be Brownian motions that are independent of each other and of W_t. The process $\sigma_{1,t}^2$ follows a two-factor stochastic volatility model: $\sigma_{1,t}^2 = v_t + \tilde{v}_t$, where

$$\begin{cases} dv_t = -0.0429(v_t - 0.1110)dt + 0.6475\sqrt{v_t}dB_{1,t}, \\ d\tilde{v}_t = -3.74(\tilde{v}_t - 0.3980)dt + 1.1656\sqrt{\tilde{v}_t}dB_{2,t}. \end{cases} \tag{5.6}$$

The process $\sigma_{2,t}^2$ is specified as a GARCH diffusion:

$$d\sigma_{2,t}^2 = -0.035(\sigma_{2,t}^2 - 0.636)dt + 0.236\sigma_{2,t}^2 dB_{3,t}. \tag{5.7}$$

The specification for the correlation process ρ_t is a GARCH diffusion for the inverse Fisher transformation of the correlation:

$$\begin{cases} \rho_t = (e^{2y_t} - 1)/(e^{2y_t} + 1), \\ dy_t = -0.03\,(y_t - 0.64)\,dt + 0.118 y_t dB_{4,t}. \end{cases} \tag{5.8}$$

In this simulation design we take the target variable to be the daily integrated correlation, which is defined as

$$IC_t \equiv \frac{QV_{12,t}}{\sqrt{QV_{11,t}}\sqrt{QV_{22,t}}}. \tag{5.9}$$

The proxy is given by the realized correlation computed using returns sampled at frequency Δ:

$$RC_t^\Delta \equiv \frac{RV_{12,t}^\Delta}{\sqrt{RV_{11,t}^\Delta}\sqrt{RV_{22,t}^\Delta}}. \tag{5.10}$$

The first forecasting model is a GARCH(1,1)–DCC(1,1) model (Engle (2002)) applied to daily returns $r_t = X_t - X_{t-1}$:

$$\text{Model C1:} \quad \begin{cases} r_{j,t} = \sigma_{j,t}\varepsilon_{j,t}, \quad \sigma_{j,t}^2 = \omega_j + \beta_j \sigma_{j,t-1}^2 + \alpha_j r_{j,t-1}^2, \quad \text{for } j = 1, 2, \\ \rho_t^\varepsilon \equiv \mathbb{E}[\varepsilon_{1,t}\varepsilon_{2,t}|\mathcal{F}_{t-1}] = \frac{Q_{12,t}}{\sqrt{Q_{11,t}Q_{22,t}}}, \quad Q_t = \begin{pmatrix} Q_{11,t} & Q_{12,t} \\ Q_{12,t} & Q_{22,t} \end{pmatrix}, \\ Q_t = \overline{Q}\,(1 - a - b) + b\,Q_{t-1} + a\,\varepsilon_{t-1}\varepsilon_{t-1}^\intercal, \quad \varepsilon_t = (\varepsilon_{1,t}, \varepsilon_{2,t}). \end{cases} \tag{5.11}$$

The forecast for IC_{t+1} is the one-day-ahead forecast of ρ^{ε}_{t+1}. The second forecasting model extends Model C1 by adding the lagged 30-minute realized correlation to the evoluation of Q_t:

$$\text{Model C2:} \quad Q_t = \overline{Q}\,(1-a-b-g) + b\,Q_{t-1} + a\,\varepsilon_{t-1}\varepsilon_{t-1}^{\intercal} + g\,RC_{t-1}^{30\text{min}}. \tag{5.12}$$

In each simulation, we set the evaluation function $f(\cdot)$ to be the loss of Model 1 less that of Model 2. We note that the competing forecasts are not engineered to be equally accurate. Therefore, for the purpose of examining size properties of our tests, the relevant null hypothesis is not that the mean-squared-error (MSE) differential is zero. Instead, we conduct one-sided EPA test with χ in (2.5) being the population MSE of Model 1 less that of Model 2.[14] The goal of this simulation study is to determine whether our feasible tests have finite-sample rejection rates similar to those of the infeasible tests (i.e., tests based on true target variables), and, moreover, whether these tests have satisfactory size properties under the "true" null hypothesis.

5.2 Results

The results for Simulations A, B and C are presented in Tables I, II and III, respectively. In the top row of each panel are the results for the infeasible tests that are implemented with the true target variable, and in the other rows are the results for feasible tests based on proxies. We consider two implementations of the Giacomini–White (GW) test: the first is based on a Newey–West estimate of the long-run variance and critical values from the standard normal distribution. The second is based on the "fixed b" asymptotics of Kiefer and Vogelsang (2005), using the Bartlett kernel. We denote these two implementations as NW and KV, respectively. The truncation lag is $3P^{1/3}$ for NW and is $0.5P$ for KV.

In Table I we observe that the GW–NW test has a tendency to over-reject, particularly for $R = 1000$, although it performs reasonably well for the longest prediction sample ($P = 2000$). In the right panels we observe that the GW–KV test has reasonable size control, although it is slightly conservative. Importantly, the use of a proxy does not lead to worse finite-sample properties than those obtained using the true target variable; the good (or bad) finite-sample properties of the feasible tests are inherited from their infeasible counterparts.

In Table II we see that both the GW–NW and the GW–KV tests have finite-sample rejection rates close to the nominal level, for all values of P and R. The feasible tests have satisfactory size properties for all but the lowest sampling frequency.

In Table III we find that the GW–NW test over-rejects, even for large sample sizes, with

[14] We compute the population MSE of each forecast by simulating a long sample with 500,000 days. Importantly, the population MSE is computed using the *true* latent target variable, whereas the feasible tests are conducted using proxies.

Proxy RV_{t+1}^{Δ}	GW–NW $P = 500$	GW–NW $P = 1000$	GW–NW $P = 2000$	GW–KV $P = 500$	GW–KV $P = 1000$	GW–KV $P = 2000$
			$R = 500$			
True $Y_{t+1}^{\dagger}$	0.04	0.11	0.06	0.01	0.02	0.02
$\Delta = 5$ sec	0.04	0.11	0.07	0.01	0.02	0.02
$\Delta = 1$ min	0.04	0.11	0.07	0.01	0.02	0.02
$\Delta = 5$ min	0.04	0.11	0.08	0.01	0.02	0.03
$\Delta = 30$ min	0.05	0.13	0.11	0.02	0.04	0.04
			$R = 1000$			
True $Y_{t+1}^{\dagger}$	0.16	0.10	0.07	0.05	0.01	0.02
$\Delta = 5$ sec	0.16	0.10	0.07	0.05	0.01	0.02
$\Delta = 1$ min	0.16	0.10	0.07	0.06	0.01	0.02
$\Delta = 5$ min	0.16	0.09	0.08	0.06	0.01	0.02
$\Delta = 30$ min	0.21	0.13	0.11	0.09	0.02	0.02

Table I: Giacomini–White test rejection frequencies for Simulation A. The nominal size is 0.05, R is the length of the estimation sample, P is the length of the prediction sample, Δ is the sampling frequency for the proxy. The left panel shows results based on a Newey–West estimate of the long run variance, the right panel shows results based on Kiefer-Vogelsang's "fixed b" asymptotics.

rejection frequencies as high as 0.27.[15] In contrast, the rejection rates of the GW–KV test are close to the nominal level, especially for $P = 1000$ and $P = 2000$. Importantly, consistent with the negligibility result, feasible tests based on proxies again have very similar properties to infeasible tests based on the actual latent target variable.

Overall, the tests generally have reasonable finite-sample size control, except for the GW–NW test applied to correlation forecasts. In all cases, the size distortions in the tests based on a proxy match those arising in the infeasible test based on the true target variable, and are not exacerbated by the use of a proxy. Hence, we conclude that the negligibility result holds well in empirically realistic scenarios. This finding is robust with respect to the choice of the truncation lag in the estimation of long-run variance. Supplemental Appendix S.C presents these robustness checks and some additional results on the disagreement between the feasible and the infeasible tests.

[15]The reason for this poor performance appears to be the relatively high persistence in the quadratic loss differentials in Simulation C. In Simulations A and B, the autocorrelations of the loss differential sequence essentially vanish at about the 50th and the 30th lag, respectively, whereas in Simulation C they remain non-negligible even at the 100th lag.

Proxy BV^{Δ}_{t+1}	GW–NW $P = 500$	GW–NW $P = 1000$	GW–NW $P = 2000$	GW–KV $P = 500$	GW–KV $P = 1000$	GW–KV $P = 2000$
			$R = 500$			
True $Y^{\dagger}_{t+1}$	0.05	0.06	0.06	0.04	0.05	0.04
$\Delta = 5$ sec	0.06	0.06	0.06	0.04	0.04	0.05
$\Delta = 1$ min	0.07	0.08	0.07	0.05	0.06	0.06
$\Delta = 5$ min	0.03	0.05	0.04	0.02	0.06	0.05
$\Delta = 30$ min	0.03	0.02	0.00	0.03	0.03	0.01
			$R = 1000$			
True $Y^{\dagger}_{t+1}$	0.03	0.04	0.04	0.03	0.04	0.05
$\Delta = 5$ sec	0.03	0.04	0.04	0.04	0.04	0.05
$\Delta = 1$ min	0.04	0.05	0.06	0.03	0.04	0.06
$\Delta = 5$ min	0.03	0.04	0.05	0.04	0.04	0.06
$\Delta = 30$ min	0.02	0.01	0.01	0.02	0.01	0.01

Table II: Giacomini–White test rejection frequencies for Simulation B. The nominal size is 0.05, R is the length of the estimation sample, P is the length of the prediction sample, Δ is the sampling frequency for the proxy. The left panel shows results based on a Newey–West estimate of the long run variance, the right panel shows results based on Kiefer-Vogelsang's "fixed b" asymptotics.

6 Application: Comparing correlation forecasts

6.1 Data and model description

We now illustrate the use of our method with an empirical application on forecasting the integrated correlation between two assets. Correlation forecasts are critical in financial decisions such as portfolio construction and risk management; see Engle (2008) for example. Standard forecast evaluation methods do not directly apply here due to the latency of the target variable, and methods that rely on an unbiased proxy for the target variable (e.g., Hansen and Lunde (2006) and Patton (2011)) cannot be used either, due to the absence of any such proxy.[16] We hence consider this an ideal example to illustrate the usefulness of the method proposed in the current paper.

[16] When based on relatively sparse sampling frequencies it *may* be considered plausible that the realized covariance matrix is finite-sample unbiased for the true quadratic covariation matrix, however as the correlation involves a ratio of the elements of this matrix, this property is lost.

Proxy RC_{t+1}^{Δ}	GW–NW $P = 500$	GW–NW $P = 1000$	GW–NW $P = 2000$	GW–KV $P = 500$	GW–KV $P = 1000$	GW–KV $P = 2000$
			$R = 500$			
True $Y_{t+1}^{\dagger}$	0.26	0.22	0.20	0.12	0.07	0.04
$\Delta = 5$ sec	0.26	0.22	0.20	0.12	0.07	0.04
$\Delta = 1$ min	0.26	0.22	0.19	0.12	0.06	0.04
$\Delta = 5$ min	0.27	0.22	0.19	0.11	0.08	0.05
$\Delta = 30$ min	0.26	0.24	0.19	0.12	0.08	0.06
			$R = 1000$			
True $Y_{t+1}^{\dagger}$	0.27	0.20	0.15	0.11	0.07	0.03
$\Delta = 5$ sec	0.28	0.20	0.15	0.11	0.07	0.03
$\Delta = 1$ min	0.27	0.20	0.15	0.11	0.07	0.03
$\Delta = 5$ min	0.27	0.20	0.16	0.10	0.07	0.03
$\Delta = 30$ min	0.25	0.20	0.13	0.12	0.06	0.03

Table III: Giacomini–White test rejection frequencies for Simulation C. The nominal size is 0.05, R is the length of the estimation sample, P is the length of the prediction sample, Δ is the sampling frequency for the proxy. The left panel shows results based on a Newey–West estimate of the long run variance, the right panel shows results based on Kiefer-Vogelsang's "fixed b" asymptotics.

Our sample consists two pairs of stocks: (i) Procter and Gamble (NYSE: PG) and General Electric (NYSE: GE) and (ii) Microsoft (NYSE: MSFT) and Apple (NASDAQ: AAPL). The sample period ranges from January 2000 to December 2010, consisting of 2,733 trading days, and we obtain our data from the TAQ database. As in Simulation C from the previous section, we take the proxy to be the realized correlation RC_t^{Δ} formed using returns with sampling interval Δ.[17] While the sampling frequency should be chosen as high as possible in the theory above, in practice we use relatively sparsely sampled data in order to reduce the effect of market microstructure effects such as the presence of a bid-ask spread and trade asynchronicity. In order to examine the robustness of our results, we consider Δ ranging from 1 minute to 130 minutes, which covers sampling intervals typically employed in empirical work.

We compare four forecasting models, all of which have the following specification for the con-

[17] For all sampling intervals we use the "subsample-and-average" estimator of Zhang, Mykland, and Aït-Sahalia (2005), with ten equally-spaced subsamples.

ditional mean and variance: for stock i, $i = 1$ or 2,

$$\begin{cases} r_{it} = \mu_i + \sigma_{it}\varepsilon_{it}, \\ \sigma_{it}^2 = \omega_i + \beta_i \sigma_{i,t-1}^2 + \alpha_i \sigma_{i,t-1}^2 \varepsilon_{i,t-1}^2 + \delta_i \sigma_{i,t-1}^2 \varepsilon_{i,t-1}^2 1_{\{\varepsilon_{i,t-1} \leq 0\}} + \gamma_i RV_{i,t-1}^{1\,\text{min}}. \end{cases} \tag{6.1}$$

That is, we assume a constant conditional mean, and a GJR-GARCH (Glosten et al. (1993)) volatility model augmented with lagged one-minute RV.

The baseline correlation model is Engle's (2002) DCC model as considered in Simulation C; see equation (5.11). The other three models are extensions of the baseline model. The first extension is the asymmetric DCC (A-DCC) model of Cappiello, Engle, and Sheppard (2006), which is designed to capture asymmetric reactions in correlation to the sign of past shocks:

$$Q_t = \overline{Q}\ (1 - a - b - d) + b\, Q_{t-1} + a\, \varepsilon_{t-1}\varepsilon_{t-1}^{\intercal} + d\, \eta_{t-1}\eta_{t-1}^{\intercal}, \quad \text{where} \quad \eta_t \equiv \varepsilon_t \circ 1_{\{\varepsilon_t \leq 0\}}. \tag{6.2}$$

The second extension (R-DCC) augments the DCC model with the 65-minute realized correlation. This extension is in the same spirit as Noureldin, Shephard, and Sheppard (2012), and is designed to exploit high-frequency information about current correlation:

$$Q_t = \overline{Q}\ (1 - a - b - g) + b\, Q_{t-1} + a\, \varepsilon_{t-1}\varepsilon_{t-1}^{\intercal} + g\, RC_{t-1}^{65\,\text{min}}. \tag{6.3}$$

The third extension (AR-DCC) encompasses both A-DCC and R-DCC with the specification

$$Q_t = \overline{Q}\ (1 - a - b - d - g) + b\, Q_{t-1} + a\, \varepsilon_{t-1}\varepsilon_{t-1}^{\intercal} + d\, \eta_{t-1}\eta_{t-1}^{\intercal} + g\, RC_{t-1}^{65\,\text{min}}. \tag{6.4}$$

We conduct pairwise comparisons of forecasts based on these four models, which include both nested and nonnested cases. We use the framework of Giacomini and White (2006), so that nested and nonnested models can be treated in a unified manner. Each one-day-ahead forecast is constructed in a rolling scheme with fixed estimation sample size $R = 1500$ and prediction sample size $P = 1233$. We use the quadratic loss function as in Simulation C.

6.2 Results

Table IV presents results for the comparison between each of the three generalized models and the baseline DCC model, using both the GW–NW and the GW–KV tests. We summarize our findings as follows. First, the A-DCC model does not improve the predictive ability over the baseline DCC model. For each stock pair, the GW–KV test reveals that the loss of the A-DCC forecast is not statistically different from that of the baseline DCC. The GW–NW test reports statistically significant outperformance of the A-DCC model relative to the DCC for some proxies. However this finding needs to be interpreted with care, as the GW–NW test was found to over-reject in finite samples in Simulation C of the previous section. Interestingly, for the MSFT–AAPL pair,

Proxy RC_{t+1}^{Δ}	GW–NW DCC vs A-DCC	GW–NW DCC vs R-DCC	GW–NW DCC vs AR-DCC	GW–KV DCC vs A-DCC	GW–KV DCC vs R-DCC	GW–KV DCC vs AR-DCC
			Panel A. PG–GE Correlation			
$\Delta = 1$ min	1.603	3.130*	2.929*	1.947	1.626	1.745
$\Delta = 5$ min	1.570	2.932*	2.724*	1.845	2.040	2.099
$\Delta = 15$ min	1.892*	2.389*	2.373*	2.047	1.945	1.962
$\Delta = 30$ min	2.177*	1.990*	2.206*	2.246	1.529	1.679
$\Delta = 65$ min	1.927*	0.838	1.089	1.642	0.828	0.947
$\Delta = 130$ min	0.805	0.835	0.688	0.850	0.830	0.655
			Panel B. MSFT–AAPL Correlation			
$\Delta = 1$ min	-0.916	2.647*	1.968*	-1.024	4.405*	3.712*
$\Delta = 5$ min	-1.394	3.566*	2.310*	-1.156	4.357*	2.234
$\Delta = 15$ min	-1.391	3.069*	1.927*	-1.195	4.279*	2.116
$\Delta = 30$ min	-1.177	3.011*	2.229*	-1.055	3.948*	2.289
$\Delta = 65$ min	-1.169	2.634*	2.071*	-1.168	3.506*	2.222
$\Delta = 130$ min	-1.068	1.825*	1.280	-1.243	3.342*	1.847

Table IV: T-statistics for out-of-sample forecast comparisons of correlation forecasting models. In the comparison of "A vs B," a positive t-statistic indicates that B outperforms A. The 95% critical values for one-sided tests of the null are 1.645 (GW–NW, in the left panel) and 2.774 (GW–KV, in the right panel). Test statistics that are greater than the critical value are marked with an asterisk.

the more general A-DCC model actually underperforms the baseline model. Second, the R-DCC model outperforms the DCC model, particularly in the MSFT–AAPL case, where the finding is highly significant and is robust to the choice of proxy. Third, the AR-DCC model also appears to outperform the DCC model. That noted, the statistical significance of the outperformance of AR-DCC depends on the testing method. In view of the over-rejection problem of the GW–NW test, we conclude with a conservative interpretation of the finding about AR-DCC: it is not significantly better than the baseline DCC.

Table V presents results from pairwise comparisons among the generalized models. Consistent with the results in Table IV, we find that the A-DCC forecast underperforms those of R-DCC and AR-DCC, and significantly so for MSFT–AAPL. The comparison between R-DCC and AR-DCC

Proxy RC_{t+1}^{Δ}	GW–NW A-DCC vs R-DCC	GW–NW A-DCC vs AR-DCC	GW–NW R-DCC vs AR-DCC	GW–KV A-DCC vs R-DCC	GW–KV A-DCC vs AR-DCC	GW–KV R-DCC vs AR-DCC
Panel A. PG–GE Correlation						
$\Delta = 1$ min	2.231*	2.718*	0.542	1.231	1.426	0.762
$\Delta = 5$ min	2.122*	2.430*	0.355	1.627	1.819	0.517
$\Delta = 15$ min	1.564	1.969*	0.888	1.470	1.703	1.000
$\Delta = 30$ min	0.936	1.561	1.282	0.881	1.271	0.486
$\Delta = 65$ min	-0.110	0.391	1.039	-0.153	0.413	0.973
$\Delta = 130$ min	0.503	0.474	-0.024	0.688	0.516	-0.031
Panel B. MSFT–AAPL Correlation						
$\Delta = 1$ min	3.110*	3.365*	-1.239	3.134*	3.657*	-1.580
$\Delta = 5$ min	4.005*	4.453*	-1.554	4.506*	6.323*	-1.586
$\Delta = 15$ min	3.616*	4.053*	-1.307	4.044*	5.449*	-1.441
$\Delta = 30$ min	3.345*	3.770*	-0.834	4.635*	7.284*	-0.882
$\Delta = 65$ min	2.999*	3.215*	-0.542	6.059*	7.868*	-0.635
$\Delta = 130$ min	2.223*	2.357*	-1.039	3.392*	5.061*	-1.582

Table V: T-statistics for out-of-sample forecast comparisons of correlation forecasting models. In the comparison of "A vs B," a positive t-statistic indicates that B outperforms A. The 95% critical values for one-sided tests of the null are 1.645 (GW–NW, in the left panel) and 2.774 (GW–KV, in the right panel). Test statistics that are greater than the critical value are marked with an asterisk.

yields mixed, but statistically insignificant, findings, across the two stock pairs.

Overall, we find that augmenting the DCC model with lagged realized correlation significantly improves its predictive ability, while adding an asymmetric term to the DCC model generally does not improve, and sometimes hurts, its forecasting performance. These findings are robust to the choice of proxy.

7 Concluding remarks

This paper proposes a simple but general framework for the problem of testing predictive ability when the target variable is unobservable. We consider an array of popular forecast evaluation methods, including Diebold and Mariano (1995), West (1996), White (2000), Romano and Wolf (2005), Giacomini and White (2006), McCracken (2007), and Hansen, Lunde, and Nason (2011), in cases where the latent target variable is replaced by a proxy computed using high-frequency (intraday) data. We provide high-level conditions under which tests based on a high-frequency proxy provide the same asymptotic properties (level and power) under null and alternative hypotheses involving the latent target variable as those involving the proxy. We then provide primitive conditions for general classes of high-frequency based estimators of volatility and jump functionals, which cover almost all existing estimators as special cases, such as realized (co)variance, truncated (co)variation, bipower variation, realized correlation, realized beta, jump power variation, realized semivariance, realized Laplace transform, realized skewness and kurtosis. In so doing, we bridge the vast literature on forecast evaluation and the burgeoning literature on high-frequency time series. The theoretical framework is structured in a way to facilitate further extensions in both directions.

The asymptotic theory reflects a simple intuition: the approximation error in the high-frequency proxy will be negligible when the proxy error is small in comparison with the magnitude of the forecast error, or more precisely, in comparison with the magnitude of the evaluation measure $f_{t+\tau}$ and its sampling variability. To the extent that ex-post measurement is easier than forecasting, this intuition, and hence our formalization, is relevant in many empirical settings. We verify that the asymptotic results perform well in three distinct, and realistically calibrated, Monte Carlo studies. Our empirical application uses these results to reveal the (pseudo) out-of-sample predictive gains from augmenting the widely-used DCC model (Engle (2002)) with high-frequency estimates of correlation.

References

Aït-Sahalia, Y., and J. Jacod (2009): "Estimating the Degree of Activity of Jumps in High Frequency Financial Data," *Annals of Statistics*, 37, 2202–2244.

——— (2012): "Analyzing the Spectrum of Asset Returns: Jump and Volatility Components in High Frequency Data," *Journal of Economic Literature*, 50, 1007–1050.

Amaya, D., P. Christoffersen, K. Jacobs, and A. Vasquez (2011): "Do Realized Skewness and Kurtosis Predict the Cross-Section of Equity Returns?," Discussion paper, University of Toronto.

Andersen, T. G., and T. Bollerslev (1998): "Answering the Skeptics: yes, standard volatility models do provide accurate forecasts," *International Economic Review*, 39, 885–905.

Andersen, T. G., T. Bollerslev, P. Christoffersen, and F. X. Diebold (2006): "Volatility and Correlation Forecasting," in *Handbook of Economic Forecasting, Volume 1*, ed. by G. Elliott, C. W. J. Granger, and A. Timmermann. Elsevier, Oxford.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003): "Modeling and Forecasting Realized Volatility," *Econometrica*, 71(2), pp. 579–625.

Andersen, T. G., T. Bollerslev, and N. Meddahi (2005): "Correcting the Errors: Volatility Forecast Evaluation Using High-Frequency Data and Realized Volatilities," *Econometrica*, 73(1), pp. 279–296.

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), pp. 817–858.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2008): "Designing realized kernels to measure the ex post variation of equity prices in the presence of noise," *Econometrica*, 76(6), 1481–1536.

Barndorff-Nielsen, O. E., S. Kinnebrouck, and N. Shephard (2010): "Measuring Downside Risk: Realised Semivariance," in *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, ed. by T. Bollerslev, J. Russell, and M. Watson, pp. 117–136. Oxford University Press.

Barndorff-Nielsen, O. E., and N. Shephard (2004a): "Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics," *Econometrica*, 72(3), pp. 885–925.

——— (2004b): "Power and bipower variation with stochastic volatility and jumps (with discussion)," *Journal of Financial Econometrics*, 2, 1–48.

Bollerslev, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31, 307–327.

Cappiello, L., R. F. Engle, and K. Sheppard (2006): "Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns," *Journal of Financial Econometrics*, 4, 537–572.

Clark, T. E., and M. W. McCracken (2005): "Evaluating Direct Multistep Forecasts," *Econometric Reviews*, 24, 369–404.

Corradi, V., and W. Distaso (2006): "Semi-Parametric Comparison of Stochastic Volatility Models Using Realized Measures," *The Review of Economic Studies*, 73(3), pp. 635–667.

Corradi, V., W. Distaso, and N. R. Swanson (2009): "Predictive Density Estimators for Daily Volatility Based on the Use of Realized Measures," *Journal of Econometrics*, 150, 119–138.

——— (2011): "Predictive Inference for Integrated Volatility," *Journal of the American Statistical Association*, 106, 1496–1512.

Corradi, V., and N. R. Swanson (2007): "Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes," *International Economic Review*, 48(1), pp. 67–109.

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

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

Diebold, F. X. (2012): "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests," Discussion paper, University of Pennsylvania.

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

Engle, R. F. (1982): "Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation," *Econometrica*, 50, 987–1008.

Engle, R. F. (2002): "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models," *Journal of Business & Economic Statistics*, 20(3), pp. 339–350.

——— (2008): *Anticipating Correlations*. Princeton University Press, Princeton, New Jersey.

Giacomini, R., and B. Rossi (2009): "Detecting and Predicting Forecast Breakdowns," *The Review of Economic Studies*, 76(2), pp. 669–705.

Giacomini, R., and H. White (2006): "Tests of conditional predictive ability," *Econometrica*, 74(6), 1545–1578.

Gonçalves, S., and R. de Jong (2003): "Consistency of the Stationary Bootstrap under Weak Moment Conditions," *Economic Letters*, 81, 273–278.

Gonçalves, S., and H. White (2002): "The Bootstrap of the Mean for Dependent Heterogeneous Arrays," *Econometric Theory*, 18(6), pp. 1367–1384.

Granger, C. (1999): "Outline of Forecast Theory Using Generalized Cost Functions," *Spanish Economic Review*, 1, 161–173.

Hansen, P. R. (2005): "A test for superior predictive ability," *Journal of Business & Economic Statistics*, 23(4), 365–380.

Hansen, P. R., and A. Lunde (2006): "Consistent Ranking of Volatility Models," *Journal of Econometrics*, 131, 97–121.

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

Hansen, P. R., and A. Timmermann (2012): "Choice of Sample Split in Out-of-Sample Forecast Evaluation," Discussion paper, European University Institute.

Huang, X., and G. T. Tauchen (2005): "The Relative Contribution of Jumps to Total Price Variance," *Journal of Financial Econometrics*, 4, 456–499.

Inoue, A., and L. Kilian (2004): "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?," *Econometric Reviews*, 23, 371–402.

Jacod, J., and P. Protter (2012): *Discretization of Processes*. Springer.

Jacod, J., and M. Rosenbaum (2012): "Quarticity and Other Functionals of Volatility: Efficient Estimation," Discussion paper, Université de Paris-6.

Jacod, J., and A. N. Shiryaev (2003): *Limit Theorems for Stochastic Processes*. Springer-Verlag, New York, second edn.

Kiefer, N. M., and T. J. Vogelsang (2005): "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory*, 21, pp. 1130–1164.

Li, J., V. Todorov, and G. Tauchen (2012): "Volatility Occupation Times," *Annals of Statistics, forthcoming.*

Mancini, C. (2001): "Disentangling the Jumps of the Diffusion in a Geometric Jumping Brownian Motion," *Giornale dell'Istituto Italiano degli Attuari*, LXIV, 19–47.

McCracken, M. W. (2000): "Robust Out-of-Sample Inference," *Journal of Econometrics*, 99, 195–223.

——— (2007): "Asymptotics for Out of Sample Tests of Granger Causality," *Journal of Econometrics*, 140, 719–752.

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

Noureldin, D., N. Shephard, and K. Sheppard (2012): "Multivariate high-frequency-based volatility (HEAVY) models," *Journal of Applied Econometrics*, 27, 907–933.

Patton, A. J. (2011): "Volatility forecast comparison using imperfect volatility proxies," *Journal of Econometrics*, 160(1), 246–256.

Patton, A. J., and A. Timmermann (2010): "Generalized Forecast Errors, A Change of Measure, and Forecast Optimality Conditions," in *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*. Oxford University Press.

Politis, D. N., and J. P. Romano (1994): "The stationary bootstrap," *Journal of the American Statistical Association*, pp. 1303–1313.

Romano, J. P., and M. Wolf (2005): "Stepwise multiple testing as formalized data snooping," *Econometrica*, 73(4), 1237–1282.

Todorov, V., and G. Tauchen (2012a): "Inverse Realized Laplace Transform for Nonparametric Volatility Density Estimation in Jump Diffusions," *Journal of the American Statistical Association*, 107, 622–635.

——— (2012b): "The Realized Laplace Transform of Volatility," *Econometrica*, 80, 1105–1127.

Vetter, M. (2010): "Limit Theorems for Bipower Variation of Semimartingales," *Stochastic Processes and their Applications*, 120, 22–38.

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

——— (2006): "Forecast Evaluation," in *Handbook of Economic Forecasting*. North Holland Press, Amsterdam.

White, H. (2000): "A reality check for data snooping," *Econometrica*, 68(5), 1097–1126.

Zhang, L. (2006): "Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach," *Bernoulli*, 12(6), 1019–1043.

Zhang, L., P. A. Mykland, and Y. Aït-Sahalia (2005): "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data," *Journal of the American Statistical Association*, 100, 1394–1411.