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: May 13, 2015

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 a high-frequency proxy 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.

[*] We wish to thank Francis Diebold, Dobrislav Dobrev, Ronald Gallant, Silvia Gonçalves, Michael McCracken, Nour Meddahi, Benoit Perron, Dacheng Xiu, conference and seminar participants at the 2013 NBER summer institute, 2013 Symposium on Econometric Theory and Applications (SETA), 2013 Econometric Society Asian Meeting, Cambridge, UCL and University of Pennsylvania for helpful comments and discussions. All errors are ours. Financial support from the NSF under grants SES-1227448 and SES-1326819 (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. In financial applications, the variables to be forecast are often risk measures, such as volatility, beta, correlation, and jump characteristics (see Andersen, Bollerslev, Christoffersen, and Diebold (2006) for a survey). 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 observable even ex post. A large literature (see West (2006) for a survey) has evolved presenting methods for (pseudo) out-of-sample inference for predictive accuracy, however existing work typically relies on the observability of the forecast target. The goal of the current paper is to provide a general methodology for extending the applicability of forecast evaluation methods to settings with unobservable forecast target variables.

Inspired by Andersen and Bollerslev (1998), we propose to evaluate competing forecasts with respect to a *proxy* of the latent target variable, with the proxy computed from high-frequency (intraday) data, in the application of forecast evaluation methods. *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 (which we label "proxy hypotheses") are not the same as those concerning the true target variable (i.e., "true hypotheses"). To fill this gap, we consider an asymptotic setting in which the proxy is constructed using data sampled from asymptotically increasing frequencies. Under this setting, the proxy hypotheses can be considered as "local" to the true hypotheses, and we provide both high-level and primitive sufficient conditions under which the moments that specify the proxy hypotheses converge sufficiently fast to their counterparts in the true hypotheses. This convergence leads to an *asymptotic negligibility* result: forecast evaluation methods using proxies have the same asymptotic size and power properties under the proxy hypotheses as under the true hypotheses. We show in three realistic Monte Carlo designs that this result works quite well in finite samples.

The strategy of using high-frequency proxies to conduct inference has proven successful in prior work on the estimation of stochastic volatility models. Bollerslev and Zhou (2002) estimate stochastic volatility models treating the realized variance as the unobserved integrated variance. Corradi and Distaso (2006) and Todorov (2009) generalize this approach by considering additional realized measures for the integrated variance using the generalized method of moments (GMM) of Hansen (1982). These authors provide theoretical justifications for this approach by providing conditions that ensure the asymptotic negligibility of the proxy error in GMM inference for stochastic volatility models. Realized measures for other volatility functionals have also been used for parametric and nonparametric estimation of stochastic volatility models: for example, Todorov, Tauchen, and Grynkiv (2011) use the realized Laplace transform of volatility (Todorov and Tauchen (2012)) for estimating parametric stochastic volatility models; Renò (2006), Kanaya and Kristensen (2010) and Bandi and Renò (2012) consider nonparametric estimation of stochastic volatility models using spot volatility estimates (Foster and Nelson (1996), Comte and Renault (1998), Kristensen (2010)).

Our asymptotic negligibility result shares the same nature as that in the important work of Corradi and Distaso (2006), among others. However, the focus of the current paper is distinct from aforementioned work in two important aspects. First, compared with (in-sample) GMM estimation, the out-of-sample forecast evaluation problem has a more complicated econometric structure. Indeed, even in the case with ex post observable forecast targets, it is well known that forecast evaluation procedures can be drastically different from each other depending on how unknown parameters in a forecast model is estimated and updated, on whether the competing forecast models are nested or nonnested, and on how critical values of tests are computed (e.g., via direct estimation or bootstrap); see, for example, Diebold and Mariano (1995), West (1996), White (2000), McCracken (2000), Hansen (2005), Giacomini and White (2006) and McCracken (2007), as well as the comprehensive review of West (2006). The apparent idiosyncrasies of these methods present a nontrivial challenge for designing a general theoretical framework for solving the latent-target problem for a broad range of evaluation methods. Second, while prior work used proxies of the volatility or its integrated functionals such as the integrated volatility and the volatility Laplace transform for estimating stochastic volatility models, forecasting applications often concern a much broader set of risk factors, such as beta, correlation, total quadratic variation, semivariance and jump variations. The broad practical scope of financial forecasting thus calls for an extensive analysis on a wide spectrum of risk measures and proxies.

The main contribution of the current paper is to address these two issues in a general and compact framework. We achieve generality by using two (sets of) high-level conditions that are designed for bridging two large literatures: forecast evaluation and high-frequency econometrics. The first set of conditions posit an abstract structure on the forecast evaluation methods; we show that these conditions are readily verified for many inference methods proposed in the existing literature, including *all* of the evaluation methods cited above, and can be readily extended to stepwise testing procedures such as Romano and Wolf (2005) and Hansen, Lunde, and Nason (2011). The second condition concerns the approximation accuracy of the high-frequency proxy relative to the latent target variable. The main technical contribution of this paper is to verify this condition under primitive conditions for general classes of high-frequency based estimators of volatility and jump risk measures in a general Itô semimartingale model for asset prices. In particular, we allow for realistic features such as leverage effect and (active) price and volatility jumps. Our results cover many existing estimators as special cases, such as realized variation (Andersen, Bollerslev, Diebold, and Labys (2003)), truncated variation (Mancini (2001)), bipower variation (Barndorff-Nielsen and Shephard (2004b)), realized covariation, beta and correlation (Barndorff-Nielsen and Shephard (2004a)), realized Laplace transform (Todorov and Tauchen (2012)), general integrated volatility functionals (Jacod and Protter (2012), Jacod and Rosenbaum (2013)), realized skewness, kurtosis and their extensions (Lepingle (1976), Jacod (2008), Amaya, Christoffersen, Jacobs, and Vasquez (2011)), and realized semivariance (Barndorff-Nielsen, Kinnebrouck, and Shephard (2010) and Patton and Sheppard (2013)). These technical results may be useful for other applications as well (e.g., Corradi and Distaso (2006) and Todorov (2009)).

The existing literature includes some work on forecast evaluation for latent target variables using proxy variables. In their seminal work, Andersen and Bollerslev (1998) advocated using realized variance as a proxy for evaluating volatility forecast models; see also Andersen, Bollerslev, Diebold, and Labys (2003) and Andersen, Bollerslev, and Meddahi (2005). A theoretical justification for this approach was proposed by Hansen and Lunde (2006) and Patton (2011), based on the availability of conditionally unbiased proxies. The unbiasedness condition considered in those papers must hold in finite samples, which is hard to verify except for certain cases: it may be plausible for realized variance in some applications, but is unlikely to hold for other realized measures (such as jump-robust measures of volatility like bipower variation, or ratios of measures like realized correlation). In contrast, our framework extends the insight of prior work with an asymptotic argument and is applicable for most known high-frequency based estimators.

With the usual caveat of asymptotic approximations in mind,[1] we note that our asymptotic negligibility result reflects a simple and robust intuition: the approximation error in the high-frequency proxy will be negligible when it is relatively small in comparison with the "intrinsic"

[1] See Section 1.3 of van der Vaart (1998) for an elaboration.

statistical uncertainty for forecast evaluation that would arise even in situations with observable targets. Since the ex post *measurement* of latent risks is generally much easier than their ex ante *prediction*, this intuition and, hence, our asymptotic formalization, should be relevant in many empirical settings. To judge the performance of the asymptotic results, we conduct three distinct, and realistically calibrated, Monte Carlo studies. The Monte Carlo evidence is supportive of our theory, and we discuss this further in Section 6.

We illustrate the usefulness of our approach in an empirical example for evaluating forecasts of the conditional correlation between stock returns. Correlation forecasting is of substantial importance in practice (Engle (2008)) but existing evaluation methods (see, e.g., Hansen and Lunde (2006), Patton (2011)) are silent on how rigorous forecast evaluation can be conducted. 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.

This paper is organized as follows. Section 2 presents the statistical setting. In Section 3 we discuss a variety of high-frequency proxies and derive bounds for their approximation accuracy. Section 4 presents the asymptotic properties of generic forecast evaluation methods using proxies, with further extensions discussed in Section 5. Monte Carlo results and an empirical application are in Sections 6 and 7, respectively. All proofs are in the appendix.

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 every $j \in \{1, \ldots, q\}$. 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. 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 setting

2.1 True hypotheses and proxy hypotheses

Let $(Y_t^\dagger)_{t\geq 1}$ be the time series to be forecast, which takes values in $\mathcal{Y} \subseteq \mathbb{R}^{\kappa_Y}$. We stress at the outset that $Y_t^\dagger$ is not observable, but a proxy Y_t is available. At time t, the forecaster uses data $\mathcal{D}_t \equiv \{D_s : 1 \leq s \leq t\}$ to form a forecast of $Y_{t+\tau}^\dagger$, where the forecast horizon $\tau \geq 1$ is fixed throughout the paper. We consider $\bar{k}$ competing sequences of forecasts of $Y_{t+\tau}^\dagger$, collected by $F_{t+\tau} \equiv (F_{1,t+\tau}, \ldots, F_{\bar{k},t+\tau})$. In practice, $F_{t+\tau}$ is often constructed from forecast models that involve some parameter β. We write $F_{t+\tau}(\beta)$ to emphasize such dependence and refer to the function $F_{t+\tau}(\cdot) : \beta \mapsto F_{t+\tau}(\beta)$ as the forecast model. Let $\hat{\beta}_t$ be an estimator constructed using (possibly a subset of) the dataset $\mathcal{D}_t$ and β^* be its "population" analogue. We do not require the forecast model to be correctly specified, so we treat β^* as a pseudo-true parameter (White (1982)).

Two types of forecasts have been considered in the literature: the actual forecast $F_{t+\tau} = F_{t+\tau}(\hat{\beta}_t)$ and the population forecast $F_{t+\tau}(\beta^*)$. This distinction is useful because a researcher may be 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 $F_{t+\tau}$, he/she can treat the actual forecast as an observable sequence (see, e.g., Diebold and Mariano (1995) and Giacomini and White (2006)) without the need for explicitly analyzing the forecast model $F_{t+\tau}(\cdot)$ and the discrepancy between $\hat{\beta}_t$ and β^*; in this case, we simply set β^* to be empty. With this convention, we can use the notation $F_{t+\tau}(\beta^*)$ also in the study of the inference for actual forecasts.

Given the target $Y_{t+\tau}^\dagger$, the performance of the competing forecasts is measured by $f_{t+\tau}^{\dagger *} \equiv f_{t+\tau}(Y_{t+\tau}^\dagger, \beta^*)$, where $f_{t+\tau}(y, \beta) \equiv f(y, F_{t+\tau}(\beta))$ for some known measurable $\mathbb{R}^{\kappa_f}$-valued function $f(\cdot)$. The function $f(\cdot)$ plays the role of an *evaluation measure.* Typically, $f(\cdot)$ computes the loss differential between competing forecasts: for example, $f(y, (F_1, F_2)) = (y - F_1)^2 - (y - F_2)^2$ in the case with quadratic loss. The proxy of $f_{t+\tau}^{\dagger *}$ is given by $f_{t+\tau}^* \equiv f_{t+\tau}(Y_{t+\tau}, \beta^*)$, which in turn can be estimated by $\hat{f}_{t+\tau} \equiv f_{t+\tau}(Y_{t+\tau}, \hat{\beta}_t)$. We then set

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

where $T + \tau$ 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.[2] In the sequel, we always assume $P \asymp T$ as $T \to \infty$ without further mention, while R may be fixed or diverge to ∞, depending on the application.

We now turn to the hypotheses of interest. We consider two classical testing problems for forecast evaluation: testing for equal predictive ability (one-sided or two-sided) and testing for superior predictive ability. Formally, we consider the following hypotheses: for some user-specified constant $\chi \in \mathbb{R}^{\kappa_f}$,

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

$$\begin{array}{c} \text{Superior} \\ \text{Predictive Ability} \\ \text{(SPA)} \end{array} \left\{ \begin{array}{ll} & H_0^{\dagger} : \mathbb{E}[\bar{f}_T^{\dagger *}] \leq \chi, \\ \text{vs.} & H_{a}^{\dagger} : \liminf_{T \to \infty} \mathbb{E}[\bar{f}_{j,T}^{\dagger *}] > \chi_j \text{ for some } j \in \{1, \ldots, \kappa_f\}, \end{array} \right. \tag{2.3}$$

where $H_{1a}^{\dagger}$ (resp. $H_{2a}^{\dagger}$) in (2.2) is the one-sided (resp. two-sided) alternative. In practice, the constant χ is often set to be zero.[3] Note that despite their assigned labels, these hypotheses can also be used to test for forecast encompassing and forecast rationality by setting the function $f(\cdot)$ properly; see, for example, West (2006).

Since the hypotheses in (2.2) and (2.3) rely on the true forecast target $Y_t^{\dagger}$, we refer to them as the *true hypotheses*. 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. Clearly, if $Y_t^{\dagger}$ were observable, these existing inference methods could be applied to test the true hypotheses by forming test statistics based on $f_{t+\tau}(Y_{t+\tau}^{\dagger}, \hat{\beta}_t)$. However, the latency of $Y_t^{\dagger}$ renders these inference methods infeasible.

[2] 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.

[3] Allowing χ to be nonzero incurs no additional cost in our derivations. This flexibility is particularly useful in the design of Monte Carlo experiment that examines the finite-sample performance of the asymptotic theory below; see Section 6 for details.

Feasible versions of these tests can be implemented with $Y^{\dagger}_{t+\tau}$ replaced by $Y_{t+\tau}$. However, the hypotheses underlying the feasible inference procedure are then *proxy hypotheses* given by

$$\begin{array}{c}\text{Proxy Equal}\\ \text{Predictive Ability}\\ \text{(PEPA)}\end{array}\left\{\begin{array}{ll} & H_0 : \mathbb{E}\left[\bar{f}^*_T\right] = \chi, \\ \text{vs.} & H_{1a} : \liminf_{T\to\infty} \mathbb{E}[\bar{f}^*_{j,T}] > \chi_j \text{ for some } j \in \{1, \ldots, \kappa_f\}, \\ \text{or} & H_{2a} : \liminf_{T\to\infty} \|\mathbb{E}[\bar{f}^*_T] - \chi\| > 0, \end{array}\right. \tag{2.4}$$

$$\begin{array}{c}\text{Proxy Superior}\\ \text{Predictive Ability}\\ \text{(PSPA)}\end{array}\left\{\begin{array}{ll} & H_0 : \mathbb{E}[\bar{f}^*_T] \le \chi, \\ \text{vs.} & H_a : \liminf_{T\to\infty} \mathbb{E}[\bar{f}^*_{j,T}] > \chi_j \text{ for some } j \in \{1, \ldots, \kappa_f\}. \end{array}\right. \tag{2.5}$$

These hypotheses are not of immediate economic relevance, because economic agents are, by assumption, interested in forecasting the true target $Y^{\dagger}_{t+\tau}$, rather than its proxy.[4]

Below, we provide conditions under which the moments that define the proxy hypotheses converge "sufficiently fast" to their equivalents under the true hypotheses, so that tests that are valid under the former are also valid under the latter. The key step in this analysis is to characterize the approximation accuracy of Y_t with respect to $Y^{\dagger}_t$. In this paper, we are mainly interested in cases where $Y^{\dagger}_t$ is a latent risk measure that takes form of a functional of the stochastic volatility and/or jumps of continuous-time asset price processes, with $Y^{\dagger}_t$ being the corresponding nonparametric estimator formed using discretely sampled data at high frequency. We now turn to the formal probabilistic setting for the high-frequency asset price data.

2.2 The underlying asset price process

In this subsection, we describe the continuous-time model for the underlying (logarithmic) asset price process X_t. Our basic assumption is that X_t is a d-dimensional Itô semimartingale defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge0}, \mathbb{P})$ with the following form

$$\begin{aligned} X_t = X_0 &+ \int_0^t b_s ds + \int_0^t \sigma_s dW_s \\ &+ \int_0^t \int_{\mathbb{R}} \delta(s,z)\, 1_{\{\|\delta(s,z)\|\le1\}} \tilde{\mu}(ds,dz) + \int_0^t \int_{\mathbb{R}} \delta(s,z)\, 1_{\{\|\delta(s,z)\|>1\}} \mu(ds,dz), \end{aligned} \tag{2.6}$$

[4] 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.

where b_t is a d-dimensional càdlàg adapted process, W_t is a d'-dimensional standard Brownian motion, σ_t is a $d \times d'$ stochastic volatility process, $\delta : \Omega \times \mathbb{R}_+ \times \mathbb{R} \mapsto \mathbb{R}^d$ is a predictable function, μ is a Poisson random measure on $\mathbb{R}_+ \times \mathbb{R}$ with compensator $\nu\left(ds, dz\right) = ds \otimes \lambda\left(dz\right)$ for some σ-finite measure λ, and $\tilde{\mu} \equiv \mu - \nu$. Itô semimartingales are widely used for modeling asset prices in financial economics and econometrics; see, for example, Duffie (2001), Singleton (2006) and Jacod and Protter (2012).

The diffusive risk and the jump risk in X_t are respectively captured by the spot covariance matrix $c_t \equiv \sigma_t \sigma_t^\intercal$ and the jump process $\Delta X_t \equiv X_t - X_{t-}$, where $X_{t-} \equiv \lim_{s \uparrow t} X_s$. In practice, these risks are often summarized as various functionals of the processes c_t and ΔX_t, which play the role of the latent forecast target $Y_t^\dagger$ in our analysis.

To simplify the discussion, we normalize the unit of time to be one day. For each day t, the process X is sampled at deterministic discrete times $t - 1 = \tau(t, 0) < \cdots < \tau\left(t, n_t\right) = t$, where n_t is the number of intraday returns. Moreover, we set $d_{t,i} = \tau(t, i) - \tau(t, i - 1)$ and denote the sampling mesh by $d_t = \max_{1 \leq i \leq n_t} d_{t,i}$. The basic assumption on the sampling scheme is that d_t should be "small" in the prediction sample, as formalized below.

Assumption S: $d_T \to 0$ and $d_T = O(n_T^{-1})$ as $T \to \infty$.

Assumption S posits that the sampling mesh and the sample span T respectively go to 0 and ∞ in a joint, rather than a sequential, way. Under this condition, we characterize the rate of convergence of various high-frequency proxies in Section 3. This sampling scheme is essentially the same as the "double asymptotic" setting considered by Corradi and Distaso (2006) and Todorov (2009), among others. Indeed, the latter amounts to setting $d_{t,i}$ to be a constant Δ, so that Assumption S posits $\Delta \to 0$ and $T \to \infty$ asymptotically. Allowing for time-varying sampling incurs no additional cost in our derivation, but is conceptually desirable in practice. As the trading activity has grown substantially over the past two decades, later samples have a much larger number of, and less noisy, intradaily observations than those in earlier samples, so it is generally more efficient to sample more frequently in later samples (Aït-Sahalia, Mykland, and Zhang (2005), Zhang, Mykland, and Aït-Sahalia (2005a), Bandi and Russell (2008)). This setting is also aligned naturally with the focal point of our approximation argument: we are interested in using the proxy $Y_{t+\tau}$ to approximate the true target $Y_{t+\tau}^\dagger$ in the prediction sample (i.e., $t \in \{R, \ldots, T\}$) for evaluation, while being agnostic about the regression sample (i.e., $t < R$).

We need the following regularity condition for the process X_t.

Assumption HF: Suppose that the following conditions hold for constants $r \in (0,2]$, $k \geq 2$ and $C > 0$.

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

$$\sigma_t = \sigma_0 + \int_0^t \tilde{b}_s ds + \int_0^t \tilde{\sigma}_s dW_s + \int_0^t \int_{\mathbb{R}} \tilde{\delta}\left(s,z\right) \tilde{\mu}\left(ds,dz\right), \tag{2.7}$$

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

(ii) For some nonnegative deterministic functions $\Gamma\left(\cdot\right)$ and $\tilde{\Gamma}(\cdot)$ on $\mathbb{R}$, we have $\|\delta\left(\omega,s,z\right)\| \leq \Gamma\left(z\right)$ and $\|\tilde{\delta}(\omega,s,z)\| \leq \tilde{\Gamma}(z)$ for all $(\omega,s,z) \in \Omega \times \mathbb{R}_+ \times \mathbb{R}$ and

$$\begin{aligned} &\int_{\mathbb{R}} (\Gamma\left(z\right)^r \wedge 1)\lambda\left(dz\right) + \int_{\mathbb{R}} \Gamma\left(z\right)^k 1_{\{\Gamma(z)>1\}}\lambda\left(dz\right) < \infty, \\ &\int_{\mathbb{R}} (\tilde{\Gamma}\left(z\right)^2 + \tilde{\Gamma}\left(z\right)^k)\lambda\left(dz\right) < \infty. \end{aligned} \tag{2.8}$$

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

$$\mathbb{E}\|b'_s\|^k + \mathbb{E}\|\sigma_s\|^k + \mathbb{E}\|\tilde{b}_s\|^k + \mathbb{E}\|\tilde{\sigma}_s\|^k \leq C. \tag{2.9}$$

Assumption HF(i) posits that the stochastic volatility process σ_t is also an Itô semimartingale. Assumption HF(ii) imposes a type of dominance condition on the random jump size for the price and the volatility. The constant r provides an upper bound for the generalized Blumenthal-Getoor index, or "activity," of jumps in X. The assumption is weaker when r is larger, in which case it is more difficult to separate jumps from the diffusive component of X_t. We do not need to restrict the activity of volatility jumps. The kth-order integrability of $\Gamma(\cdot)$ and $\tilde{\Gamma}\left(\cdot\right)$ places restrictions on jump tails and it facilitates the derivation of bounds via sufficiently high moments. Assumption HF(iii) imposes integrability conditions that serve the same purpose.[5]

3 High-frequency proxies and their accuracy

In this section, we introduce proxies Y_t for various risk measures $Y_t^{\dagger}$ and provide convergence rate results under the L_p norm. Sections 3.1–3.3 consider three general classes of proxies and Section 3.4 considers some additional important examples. We show that $\|Y_t - Y_t^{\dagger}\|_p \leq K d_t^{\theta}$ for constants

[5] Technically speaking, this condition could be further relaxed so that the moments may be moderately explosive (see, e.g., Kanaya and Kristensen (2010)).

K and θ, where the rate θ varies across these settings. These results are the main technical contribution of the current paper and are essential for interpreting the regularity conditions and, ultimately, the asymptotic negligibility result in Section 4. We stress that, unlike the existing convergence rate results under the fixed-T setting (see, e.g., Jacod and Protter (2012)), we consider rate results that are valid in the large-T setting, which demands different conditions and proofs. Below, for each $t \geq 1$ and $i \geq 1$, we denote the ith return of X in day t by $\Delta_{t,i}X$, that is, $\Delta_{t,i}X \equiv X_{\tau(t,i)} - X_{\tau(t,i-1)}$. We suppose that the sampling mesh sequence $(d_t)_{t\geq 0}$ satisfies Assumption S throughout this section.

3.1 Generalized realized variations for continuous processes

We start with the basic setting with X continuous; the continuity condition will be relaxed in later subsections. We consider the following general class of estimators: for any measurable function $g : \mathbb{R}^d \mapsto \mathbb{R}$, we set

$$\widehat{\mathcal{I}}_t\,(g) \equiv \sum_{i=1}^{n_t} g(\Delta_{t,i}X/d_{t,i}^{1/2})d_{t,i}.$$

We also associate g with the following function: for any $d \times d$ positive semidefinite matrix A, we set $\rho\,(A; g) \equiv \mathbb{E}\,[g\,(U)]$ for $U \sim \mathcal{N}\,(0, A)$, provided that the expectation is well-defined. Theorem 3.1 below provides a bound for the approximation error between the proxy $Y_t = \widehat{\mathcal{I}}_t(g)$ and the target variable $Y_t^\dagger = \mathcal{I}_t(g) \equiv \int_{t-1}^t \rho\,(c_s; g)\,ds$.

In many applications, the function $\rho\,(\,\cdot\,; g)$ and, hence, $\mathcal{I}_t(g)$ can be expressed in closed form. For example, in the scalar case (i.e., $d = 1$), if we take $g\,(x) = |x|^a/m_a$ for some $a \geq 2$, where m_a is the ath absolute moment of a standard normal variable, then $\mathcal{I}_t(g) = \int_{t-1}^t c_s^{a/2}ds$; the integrated variance is a special case with $a = 2$. Another univariate example is to take $g(x) = \cos(\sqrt{2u}x)$, $u > 0$, 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 (2012)) and $\mathcal{I}_t(g)$ is the Laplace transform of the volatility occupation density which captures the distributional information of volatility. A simple bivariate example is $g(x_1, x_2) = x_1x_2$, 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; see Barndorff-Nielsen and Shephard (2004a).

Theorem 3.1. *Let $p \in [1, 2)$ and $C > 0$ be constants. Suppose (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 $\|\widehat{\mathcal{I}}_t(g) - \mathcal{I}_t(g)\|_p \leq Kd_t^{1/2}$ for some constant $K > 0$ and all t.*

3.2 Jump-robust proxies for integrated volatility functionals

We now turn to a general setting in which X_t may have jumps. In this subsection, we consider jump-robust proxies for risk measures with the form $\mathcal{I}_t^\star(g) = \int_{t-1}^t g(c_s)ds$, where $g : \mathbb{R}^{d\times d} \mapsto \mathbb{R}$ is a twice continuously differentiable function with at most polynomial growth. This class of risk factors is quite general: integrated variance and covariance, integrated quarticity, and volatility Laplace and Fourier transforms are special cases.[6]

In order to construct a jump-robust proxy for $\mathcal{I}_t^\star(g)$, we first nonparametrically recover the spot covariance process by using a spot truncated covariation estimator given by[7]

$$\hat{c}_{\tau(t,i)} = \frac{1}{k_t}\sum_{j=1}^{k_t} d_{t,i+j}^{-1}\Delta_{t,i+j}X\Delta_{t,i+j}X^\intercal 1_{\{\|\Delta_{t,i+j}X\|\le\bar{\alpha}d_{t,i+j}^{\varpi}\}}, \tag{3.1}$$

where $\bar{\alpha} > 0$ and $\varpi \in (0, 1/2)$ are constant tuning parameters, and k_t is an integer that specifies the local window for the spot covariance estimation and may vary across days. We consider the sample analogue of $\mathcal{I}_t^\star(g)$ as its proxy, that is, $\widehat{\mathcal{I}}_t^\star(g) = \sum_{i=0}^{n_t-k_t} g(\hat{c}_{\tau(t,i)})d_{t,i}$.

Theorem 3.2. *Let $q \geq 2$, $p \in [1, 2)$ and $C > 0$ be constants. Suppose (i) g is twice continuously differentiable and $\|\partial_x^j g(x)\| \leq C(1 + \|x\|^{q-j})$ for $j \in \{0, 1, 2\}$; (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 that σ_t is continuous. We also set $\theta_2 = \min\{1 - \varpi r + q(2\varpi - 1), 1/r - 1/2\}$. Then $\|\widehat{\mathcal{I}}_t^\star(g) - \mathcal{I}_t^\star(g)\|_p \leq K d_t^{\theta_1\wedge\theta_2}$ for some constant K and all t.*

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.

[6] Jump-robust estimators for the integrated volatility, such as the bipower and the tripower variations, were studied by Corradi and Distaso (2006) and Todorov (2009).

[7] Spot variance estimators can be dated back to Foster and Nelson (1996) and Comte and Renault (1998); also see Kristensen (2010) and references therein. The truncation technique was proposed by Mancini (2001) for the estimation of integrated variance. The spot truncated covariation estimator appeared in Chapter 9 of Jacod and Protter (2012), although they have been considered as auxiliary results in other contexts (see, e.g., Aït-Sahalia and Jacod (2009)).

(ii) Jacod and Rosenbaum (2013) characterize the limit distribution of $\widehat{\mathcal{I}}_t^\star(g)$ under the in-fill asymptotic setting with a fixed time 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 a setting with $d_T \to 0$ and $T \to \infty$. Our results also cover the case with active jumps, that is, the setting with $r \geq 1$.

3.3 Functionals of price jumps

In this subsection, we consider jump risk measures. The target variable of interest takes 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 $\widehat{\mathcal{J}}_t\left(g\right) \equiv \sum_{i=1}^{n_t} g\left(\Delta_{t,i}X\right)$. 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 ($g(x_1, x_2) = x_1^2 x_2^2$). See Amaya, Christoffersen, Jacobs, and Vasquez (2011) for applications of these risk factors.

Theorem 3.3. *Let $p \in [1, 2)$ and $C > 0$ be constants. Suppose (i) g is twice continuously differentiable; (ii) for some $q_2 \geq q_1 \geq 3$, 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 $\|\widehat{\mathcal{J}}_t\left(g\right) - \mathcal{J}_t(g)\|_p \leq K d_t^{1/2}$ for some constant K and all t.*

Comment. 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.

3.4 Additional special examples

We now consider a few special examples which are not covered by Theorems 3.1–3.2. In the first example, the true target is the daily quadratic covariation matrix QV_t of the process X, that is, $QV_t \equiv \int_{t-1}^t c_s ds + \sum_{t-1<s\leq t} \Delta X_s \Delta X_s^\intercal$. The associated proxy is the realized covariation matrix

$$RV_t \equiv \sum_{i=1}^{n_t} \Delta_{t,i} X \Delta_{t,i} X^\intercal. \tag{3.2}$$

Theorem 3.4. *Let $p \in [1, 2)$. Suppose Assumption HF with $k \geq \max\{2p/(2-p), 4\}$. Then $\|RV_t - QV_t\|_p \leq K d_t^{1/2}$ for some K and all t.*

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

$$BV_t = \frac{n_t}{n_t - 1}\frac{\pi}{2}\sum_{i=1}^{n_t - 1} |d_{t,i}^{-1/2}\Delta_{t,i}X||d_{t,i+1}^{-1/2}\Delta_{t,i+1}X|d_{t,i}. \tag{3.3}$$

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

Theorem 3.5. *Let p and p' be constants such that $1 \leq p < p' \leq 2$. Suppose that Assumption HF holds with $d = 1$ and $k \geq \max\{pp'/(p'-p), 4\}$. We have, for some K and all t, (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 in the case with jumps. The rate is slower than that in the continuous case. The constant p' arises as a technical device in our proofs and should be chosen close to p so that the bound in part (a) is sharper. We note that, the rate in part (a) is sharper when r is smaller. 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.[8]

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 $SV_t^{\pm} = \frac{1}{2}\int_{t-1}^{t} c_s ds + \sum_{t-1<s\leq t}\{\Delta X_s\}_{\pm}^2$.

Theorem 3.6. *Let p and p' be constants such that $1 \leq p < p' \leq 2$. Suppose that Assumption HF holds with $d = 1$, $r \in (0,1]$ and $k \geq \max\{pp'/(p'-p), 4\}$. Then for some K and all t, (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,

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

and Shephard (2010) under the fixed-span 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.

4 Forecast evaluation methods with high-frequency proxies

This section presents the asymptotic properties of the feasible evaluation methods using proxies. In Section 4.1 we introduce high-level conditions that link many apparently distinct tests of predictive accuracy into a unified framework. In Section 4.2, we discuss regularity conditions for the asymptotic negligibility of the high-frequency proxy errors; these conditions are motivated by the convergence rate results in Section 3. Section 4.3 presents asymptotic properties of the feasible evaluation procedures.

4.1 Conditions on evaluation methods for the proxy hypotheses

In this subsection, we introduce an abstract econometric structure that is common to most forecast evaluation procedures with an observable forecast target, the role of which is played by the proxy Y_t in the setting of the current paper. These conditions speak to the proxy hypotheses PEPA and PSPA, but not the true hypotheses; conditions in Section 4.2 below fill this gap.

We consider a test statistic of the form

$$\varphi_T \equiv \varphi(a_T(\bar{f}_T - \chi), a'_T S_T) \tag{4.1}$$

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.[9] 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 4.4 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 4.2 below); $\mathcal{S}$ is then the space of positive definite matrices.

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

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.

Assumption A1: $(a_T(\bar{f}_T - \mathbb{E}[\bar{f}_T^*]), a'_T S_T) \xrightarrow{d} (\xi, S)$ for some deterministic sequences $a_T \to \infty$ 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.

Assumption A1 mainly posits that $\bar{f}_T$ is centered at $\mathbb{E}[\bar{f}_T^*]$ with a well-behaved asymptotic distribution. Since $\mathbb{E}[\bar{f}_T^*]$ characterizes the proxy hypotheses (recall (2.4) and (2.5)), Assumption A1 concerns an evaluation problem with the observed proxy instead of the latent true target. This assumption covers many existing methods that involve observable forecast targets, as we now illustrate in detail.

Example 4.1: 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, $f(Y_{t+\tau}, (F_{1,t+\tau}, F_{2,t+\tau})) = L(Y_{t+\tau}, F_{1,t+\tau}) - L(Y_{t+\tau}, F_{2,t+\tau})$ for some loss function $L(\cdot,\cdot)$. Moreover, one can set β^* to be empty and treat each actual forecast as an observed sequence, so $\bar{f}_T = \bar{f}_T^*$. 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 long-run variance denoted by Σ. We then set $S = \Sigma$ and $a'_T \equiv 1$, and let S_T be a heteroskedasticity and autocorrelation consistent (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 4.2: Consider the same setting as in Example 4.1, 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 4.3: West (1996) considers inference for 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 $f_{t+\tau}(\cdot)$ with respect to β and concerns $\hat{\beta}_t$ in the recursive scheme. Similar results allowing for a nondifferentiable $f_{t+\tau}(\cdot)$ function can be found in McCracken (2000). Giacomini and Rossi (2009) generalize West's theory to settings without covariance stationarity. Assumption A1 can be verified similarly in these more general settings.

Example 4.4: McCracken (2007) considers inference on nested forecast models allowing for recursive, rolling and fixed estimation schemes, all with $R \to \infty$. The evaluation measure $\hat{f}_{t+\tau}$ 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 $\widehat{\Omega}_T = P^{-1}\sum_{t=R}^{T}(\hat{f}_{t+\tau} - \bar{f}_T)^2$ 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 extensions of McCracken (2007); see, for example, Inoue and Kilian (2004), Clark and McCracken (2005) and Hansen and Timmermann (2012).

Example 4.5: 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 4.3, 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 used in forecast evaluation: 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 $z_{1-\alpha}$. This assumption is slightly stronger than what we actually need. Indeed, we only need the convergence to hold under the null hypothesis, while, under the alternative, we only need the sequence $z_{T,1-\alpha}$ to be tight.

Below, we discuss examples for which Assumption A3 can be verified.

Example 4.6: 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 4.2 or 4.4, 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.[10] Assumption A3 for these examples can be verified by simply taking $z_{T,1-\alpha}$ as the known quantile of the limit distribution.

Example 4.7: 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 covariance matrix 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

[10] One caveat is that the OOS-t statistic in McCracken (2007) is asymptotically pivotal only under the somewhat restrictive condition that the forecast errors form a conditionally homoskedastic martingale difference sequence. In the presence of conditional heteroskedasticity or serial correlation in the forecast errors, the 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)).

$z_{1-\alpha}$.[11] 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 (i) $\varphi(u,s) \leq \varphi(u',s)$ whenever $u \leq u'$, where $u, u' \in \mathbb{R}^{\kappa_f}$; (ii) $\varphi(u,\tilde{s}) \to \infty$ 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(i) 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(ii) 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.

4.2 Conditions for the asymptotic negligibility of proxy errors

In this subsection, we discuss the key condition that fills the gap between the proxy hypotheses and the true hypotheses. That is, their difference $\mathbb{E}[\bar{f}_T^*] - \mathbb{E}[\bar{f}_T^{\dagger *}]$ goes to zero slightly faster than the sampling variability in the feasible test, so that the difference is negligible for asymptotic inference. The formal condition is given by Assumption C below, where a_T is given by Assumption A1.

Assumption C: $a_T(\mathbb{E}[\bar{f}_T^*] - \mathbb{E}[\bar{f}_T^{\dagger *}]) \to 0$.

Assumption C is closely related to the convergence rate results in Section 3. We have shown that for $p \in [1,2)$ and $\theta > 0$,

$$\|Y_t - Y_t^\dagger\|_p \leq K d_t^\theta, \tag{4.2}$$

[11] 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.

for various risk measures and proxies, where θ indicates the approximation accuracy. Given these technical results, Assumption C mainly requires that the sequence $(d_t)_{t\geq 1}$ of sampling meshes goes to zero sufficiently fast relative to $T \to \infty$, provided that the evaluation measure $f(\cdot)$ is smooth in the target variable. We illustrate the verification of Assumption C by examples, where K denotes a constant that may vary from line to line. We also remind the reader that $P \asymp T$ is a maintained assumption and, for the known examples in Section 4.1, $a_T = T^{\iota}$ for $\iota = 1/2$ or 1.

Example 4.8: Consider a forecast comparison setting with the evaluation measure being the loss differential of two competing forecasts, that is, $f(Y_{t+\tau}, (F_{1,t+\tau}, F_{2,t+\tau})) = L(Y_{t+\tau} - F_{1,t+\tau}) - L(Y_{t+\tau} - F_{2,t+\tau})$, where $L(\cdot)$ is a loss function. If $L(\cdot)$ is Lipschitz (e.g. Lin-Lin loss), then $|f(Y_{t+\tau}, (F_{1,t+\tau}, F_{2,t+\tau})) - f(Y^{\dagger}_{t+\tau}, (F_{1,t+\tau}, F_{2,t+\tau}))| \leq K\|Y^{\dagger}_{t+\tau} - Y_{t+\tau}\|$. Under (4.2),

$$|a_T(\mathbb{E}[\bar{f}^{*}_{T}] - \mathbb{E}[\bar{f}^{\dagger *}_{T}])| \leq K a_T P^{-1} \sum_{t=R}^{T} d_t^{\theta} \leq K T^{\iota - 1} \sum_{t=1}^{T} d_t^{\theta}. \tag{4.3}$$

By Kronecker's Lemma, the bound in (4.3) goes to zero if $\sum_{t=1}^{T} t^{\iota-1} d_t^{\theta} < \infty$. This summability condition implicitly restricts the rate at which $d_T \to 0$. It is satisfied if $d_T = O(T^{-\iota/\theta}(\log T)^{-1/\theta - \eta})$ for some $\eta > 0$ that is arbitrarily small but fixed; see Theorem 2.31 in Davidson (1994). More specifically, if we have $\iota = 1/2$ and $\theta = 1/2$ as in many basic cases, the sufficient condition for Assumption C amounts to letting the high-frequency mesh go to zero slighly faster than T^{-1}.

Example 4.9: Non-Lipschitz loss functions can also be accommodated. Consider the same setting as in Example 4.8 but with $L(\cdot)$ being the quadratic loss (i.e., $L(x) = x^2$). We have $f(Y_{t+\tau}, (F_{1,t+\tau}, F_{2,t+\tau})) - f(Y^{\dagger}_{t+\tau}, (F_{1,t+\tau}, F_{2,t+\tau})) = 2(Y_{t+\tau} - Y^{\dagger}_{t+\tau})(F_{2,t+\tau} - F_{1,t+\tau})$. Suppose $\sup_{t\geq 1}(\|F_{1,t+\tau}\|_q + \|F_{2,t+\tau}\|_q) < \infty$ for $q = p/(p-1)$.[12] By (4.2) and Hölder's inequality, we have (4.3). As in Example 4.8, Assumption C is implied by the same conditions on $(d_t)_{t\geq 1}$ discussed there.

As shown in Examples 4.8 and 4.9, the bound (4.2) facilitates the interpretation of Assumption C as a condition on the relative rates of the high-frequency mesh and the time span. This condition is weaker when θ is higher (i.e., more accurate proxies) and when a_T diverges more slowly (i.e., larger sampling uncertainty in the out-of-sample testing problem). Besides the examples in Section 3, additional results known in the literature can also be invoked for establishing (4.2). For example,

[12] Uniform boundedness on moments are commonly used for deriving asymptotic results for heterogeneous data; see, for example, White (2001). This condition is trivially satisfied if the forecasts $F_{1,t}$ and $F_{2,t}$ are bounded (e.g. forecasts for correlations).

Corradi and Distaso (2006) and Corradi, Distaso, and Swanson (2009, 2011) consider various proxies for the integrated variance which are robust to certain types of microstructure noise. These authors show that the two-scale realized variance (Zhang, Mykland, and Aït-Sahalia (2005b)), the multi-scale realized variance (Zhang (2006)) and the realized kernel (Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008)) satisfy (4.2) with $\theta = 1/6, 1/4$ and $1/4$, respectively.[13]

Finally, we note that the convergence rate condition (4.2) is stable under Hölder-continuous transforms. Indeed, if $H(\cdot)$ is a Hölder-continuous function with exponent $\hbar \in (0,1]$, then it is easy to see that $\|H(Y_t) - H(Y_t^\dagger)\|_p \leq K\|Y_t - Y_t^\dagger\|_{\hbar p}^{\hbar}$. Hence, applying (4.2) under the $L_{\hbar p}$-norm further implies $\|H(Y_t) - H(Y_t^\dagger)\|_p \leq K d_t^{\hbar\theta}$. Existing convergence rates results such as those derived in Section 3 can be "mixed and matched" to derive similar results for transformed proxies. A concrete example is given below.

Example 4.10: Consider correlation forecasting for a bivariate asset price process $X_t = (X_{1t}, X_{2t})$. Let $Y_t^\dagger = \int_{t-1}^{t} c_s ds$ be the integrated covariance matrix and Y_t be a proxy of it (see, e.g., Theorems 3.1 and 3.2). For some (small) constant $\underline{c} > 0$, we consider a transform $H(\cdot;\underline{c})$ given by

$$H(Y;\underline{c}) = \frac{Y_{12}}{\sqrt{(Y_{11} \vee \underline{c})\,(Y_{22} \vee \underline{c})}}.$$

We consider $H(Y_t^\dagger;\underline{c})$ as a measure for the correlation. Note that the integrated correlation considered by Barndorff-Nielsen and Shephard (2004a) is $H(Y_t^\dagger;0)$. Our motivation of considering $H(Y_t^\dagger;\underline{c})$ for $\underline{c} > 0$ is that the transform $H(\cdot;\underline{c})$ is Lipschitz continuous, so that (4.2) implies $\|H(Y_t;\underline{c}) - H(Y_t^\dagger;\underline{c})\|_p \leq K d_t^{\theta}$ without further assumptions. This slight modification of the integrated correlation is hardly consequential in practical terms. Indeed, if we assume that the spot volatility processes $\sqrt{c_{11,t}}$ and $\sqrt{c_{22,t}}$ are bounded below by $\sqrt{\underline{c}}$ (say, 0.1% in annualized terms), then $H(Y_t^\dagger;\underline{c}) = H(Y_t^\dagger;0)$ identically.

4.3 Asymptotic properties of the feasible inference procedure

Under the conditions discussed in Sections 4.1 and 4.2, Proposition 4.1 shows that the feasible test ϕ_T is valid under the true hypotheses.

Proposition 4.1. *The following statements hold under Assumptions A1–A3 and C.*

[13] See Propositions 3–6 in Corradi, Distaso, and Swanson (2009) and Lemma 1 in Corradi, Distaso, and Swanson (2011) for more details.

(a) Under the EPA setting (2.2), $\mathbb{E}\phi_T \to \alpha$ *under* $H_0^\dagger$*. If Assumption B1(ii) (resp. B2) holds in addition, we have* $\mathbb{E}\phi_T \to 1$ *under* $H_{1a}^\dagger$ *(resp.* $H_{2a}^\dagger$*).*

(b) Under the SPA setting (2.3) and Assumption B1, we have $\limsup_{T\to\infty} \mathbb{E}\phi_T \leq \alpha$ *under* $H_0^\dagger$ *and* $\mathbb{E}\phi_T \to 1$ *under* $H_a^\dagger$*.*

It can be shown that the test ϕ_T satisfies the same asymptotic level and power properties under the proxy hypotheses, without requiring Assumption C. Assumption C is needed for deriving asymptotic properties of ϕ_T under the true hypotheses. In particular, Proposition 4.1 shows that the level and power properties of the test are the same for the true and the proxy hypotheses. In this sense, the proxy error is negligible for the asymptotic inference about preditive accuracy.

The result established in Proposition 4.1 is a form of *weak* negligibility for the proxy error, 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 one would use if one could observe the true forecast target. We may consider the difference between the proxy and the target negligible in a *strong* sense if $\mathbb{P}(\phi_T = \phi_T^\dagger) \to 1$.[14] It is obvious that strong negligibility implies weak negligibility. While the strong negligibility may seem to be a reasonable result to pursue, we note 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.

Similar to our negligibility result, West (1996) defines cases exhibiting "asymptotic irrelevance" as those in which valid inference about predictive accuracy can be made while ignoring the presence of parameter estimation error $\hat{\beta}_t - \beta^*$. 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^*. 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, nor in that of Corradi and Distaso (2006), among others. In Section 6, we show that this negligibility result provides excellent finite-sample approximation in three realistic Monte Carlo designs.

[14] Since the tests take values in $\{0, 1\}$, $\mathbb{P}(\phi_T = \phi_T^\dagger) \to 1$ is equivalent to $\phi_T - \phi_T^\dagger \xrightarrow{\mathbb{P}} 0$.

5 Extensions: additional forecast evaluation methods

In this section we discuss several extensions of our baseline result (Proposition 4.1). 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 presented above through known, but distinct, examples. The stepwise procedures (Romano and Wolf (2005), Hansen, Lunde, and Nason (2011)) 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 Supplement to this paper.

5.1 Tests for instrumented conditional moment equalities

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

$$H_0^\dagger : \mathbb{E}[g(Y_{t+\tau}^\dagger, F_{t+\tau}(\beta^*))|\mathcal{H}_t] = 0, \quad \text{all } t \geq 0, \tag{5.1}$$

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 (5.1) implies the following unconditional moment equality:

$$H_{0,h}^\dagger : \mathbb{E}[g(Y_{t+\tau}^\dagger, F_{t+\tau}(\beta^*)) \otimes h_t] = 0, \quad \text{all } t \geq 0. \tag{5.2}$$

We cast (5.2) in the setting of Section 2 by setting $f_{t+\tau}(y,\beta) \equiv g(y, F_{t+\tau}(\beta)) \otimes h_t$. Then the theory in Section 4 can be applied without further change. In particular, Proposition 4.1 suggests that the two-sided PEPA test (with $\chi = 0$) using the proxy has a valid asymptotic level under $H_0^\dagger$ and is consistent against the alternative

$$H_{2a,h}^\dagger : \liminf_{T\to\infty} \|\mathbb{E}[g(Y_{t+\tau}^\dagger, F_{t+\tau}(\beta^*)) \otimes h_t]\| > 0. \tag{5.3}$$

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, respectively.

Example 5.1: 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 (5.1) 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 (5.3) and the power generally depends on the choice of h_t.

Example 5.2: 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 (5.1) 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.

5.2 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 $f^{\dagger*}_{j,t+\tau}=L(Y^{\dagger}_{t+\tau},F_{0,t+\tau}(\beta^*))-L(Y^{\dagger}_{t+\tau},F_{j,t+\tau}(\beta^*))$, $1\leq j\leq\bar{k}$, 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

$$\text{Multiple SPA}\left\{\begin{array}{l} H^{\dagger}_{j,0}:\mathbb{E}[f^{\dagger*}_{j,t+\tau}]\leq 0\text{ for all }t\geq 1,\\ H^{\dagger}_{j,a}:\liminf_{T\to\infty}\mathbb{E}[\bar{f}^{\dagger*}_{j,T}]>0,\end{array}\right.\qquad 1\leq j\leq\bar{k}. \tag{5.4}$$

These hypotheses concern the true target variable and are stated in a way that allows 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 condition on proxy accuracy (Assumption C), we can show that the StepM procedure, when applied to the proxy, asymptotically controls the FWE for the hypotheses (5.4) that concern the latent target. The details are in Supplemental Appendix B.1.

5.3 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

$$\liminf_{T \to \infty} \left(\overline{\mathcal{M}}^{\dagger} \subseteq \widehat{\mathcal{M}}_{T,1-\alpha} \right) \geq 1 - \alpha, \quad \mathbb{P}\left(\widehat{\mathcal{M}}_{T,1-\alpha} \cap \underline{\mathcal{M}}^{\dagger} = \emptyset \right) \to 1. \tag{5.5}$$

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 C, we can show that this feasible version achieves (5.5). The details are in Supplemental Appendix B.2.

6 Monte Carlo analysis

6.1 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, equation (2.2), 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 $R \in \{250, 500, 1000\}$ 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. As on the New York stock exchange, each day is assumed to contain 6.5 trading hours. There are 1000 Monte Carlo trials in each experiment and 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:

$$\begin{cases} dX_t = 0.0314dt + \sigma_t(-0.5760dW_{1,t} + \sqrt{1-0.5760^2}dW_{2,t}) + dJ_t, \\ d\log\sigma_t^2 = -0.0136(0.8382 + \log\sigma_t^2)dt + 0.1148dW_{1,t}, \end{cases} \tag{6.1}$$

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 $\log IV_t \equiv \log\int_{t-1}^{t}\sigma_s^2 ds$ and the proxy is $\log RV_t^{\Delta}$, where RV_t^{Δ} is defined by (3.2) for data sampled at $\Delta =$ 5 seconds, 1 minute, 5 minutes, or 30 minutes.

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

$$\text{Model A1:} \quad \begin{cases} r_t = X_t - X_{t-1} = \sigma_t\varepsilon_t, \quad \varepsilon_t|\mathcal{F}_{t-1} \sim \mathcal{N}(0,1), \\ \sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha r_{t-1}^2. \end{cases} \tag{6.2}$$

The second model is a heterogeneous autoregressive (HAR) model (Corsi (2009)) for $RV_t^{5\min}$

estimated via ordinary least squares:

$$\text{Model A2:}\quad \begin{cases} RV_t^{5\text{min}} &= \beta_0 + \beta_1 RV_{t-1}^{5\text{min}} + \beta_2 \sum_{k=1}^{5} RV_{t-k}^{5\text{min}} \\ & \quad + \beta_3 \sum_{k=1}^{22} RV_{t-k}^{5\text{min}} + e_t. \end{cases} \tag{6.3}$$

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 (6.1) 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}, where BV_t^{Δ} is defined by (3.3) for data sampled with observation mesh Δ.

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 $BV_t^{5\text{min}}$:

$$\text{Model B1:}\quad BV_t^{5\text{min}} = BV_{t-1}^{5\text{min}} + \varepsilon_t, \quad \text{where} \quad \mathbb{E}\left[\varepsilon_t | \mathcal{F}_{t-1}\right] = 0. \tag{6.4}$$

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

$$\text{Model B2:}\quad \begin{cases} BV_t^{1\text{min}} &= \beta_0 + \beta_1 BV_{t-1}^{1\text{min}} + \beta_2 \sum_{k=1}^{5} BV_{t-k}^{1\text{min}} \\ & \quad + \beta_3 \sum_{k=1}^{22} BV_{t-k}^{1\text{min}} + e_t. \end{cases} \tag{6.5}$$

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 7 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

$$\begin{cases} dv_t = -0.0429(v_t - 0.1110)dt + 0.2788\sqrt{v_t}dB_{1,t}, \\ d\tilde{v}_t = -3.7400(\tilde{v}_t - 0.3980)dt + 2.6028\sqrt{\tilde{v}_t}dB_{2,t}. \end{cases} \tag{6.6}$$

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

$$d\sigma^2_{2,t} = -0.0350(\sigma^2_{2,t} - 0.6360)dt + 0.2360\sigma^2_{2,t}dB_{3,t}. \tag{6.7}$$

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

$$\begin{cases} \rho_t = (e^{2y_t} - 1)/(e^{2y_t} + 1), \\ dy_t = -0.0300\,(y_t - 0.6400)\,dt + 0.1180 y_t dB_{4,t}. \end{cases} \tag{6.8}$$

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

$$IC_t \equiv \frac{QV_{12,t}}{\sqrt{QV_{11,t}}\sqrt{QV_{22,t}}}. \tag{6.9}$$

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

$$RC^{\Delta}_t \equiv \frac{RV^{\Delta}_{12,t}}{\sqrt{RV^{\Delta}_{11,t}}\sqrt{RV^{\Delta}_{22,t}}}. \tag{6.10}$$

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

$$\text{Model C1:} \quad \begin{cases} r_{j,t} = \sigma_{j,t}\varepsilon_{j,t}, \quad \sigma^2_{j,t} = \omega_j + \beta_j \sigma^2_{j,t-1} + \alpha_j r^2_{j,t-1}, \quad \text{for } j = 1, 2, \\ \rho^{\varepsilon}_t \equiv \mathbb{E}[\varepsilon_{1,t}\varepsilon_{2,t}|\mathcal{F}_{t-1}] = \frac{Q_{12,t}}{\sqrt{Q_{11,t}Q_{22,t}}}, \quad Q_t = \begin{pmatrix} Q_{11,t} & Q_{12,t} \\ \bullet & Q_{22,t} \end{pmatrix}, \\ Q_t = \overline{Q}\,(1 - a - b) + b\,Q_{t-1} + a\,\varepsilon_{t-1}\varepsilon^{\intercal}_{t-1}, \quad \varepsilon_t = (\varepsilon_{1,t}, \varepsilon_{2,t}). \end{cases} \tag{6.11}$$

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 evolution of Q_t:

$$\text{Model C2:} \quad Q_t = \overline{Q}\,(1 - a - b - g) + b\,Q_{t-1} + a\,\varepsilon_{t-1}\varepsilon^{\intercal}_{t-1} + g\,RC^{30\text{min}}_{t-1}. \tag{6.12}$$

In each simulation, we set the evaluation function $f(\cdot)$ to be the loss of Model 1 less that of Model 2 and conduct the one-sided EPA test (see equation (2.2)). We note that the competing forecasts are not engineered to have the same mean-squared error (MSE). Therefore, for the purpose of examining size properties of the tests, the hypotheses to be imposed are those in (2.2) with χ being the population MSE of Model 1 less that of Model 2. We remind the reader that the population MSE is computed using the *true* latent target variable, whereas the feasible tests are conducted using proxies. The goal of this simulation study is to determine whether our feasible

	GW–NW			GW–KV		
Proxy RV_{t+1}^{Δ}	$P = 500$	$P = 1000$	$P = 2000$	$P = 500$	$P = 1000$	$P = 2000$
	$R = 250$					
True $Y_{t+1}^{\dagger}$	0.08	0.07	0.07	0.01	0.01	0.01
$\Delta = 5$ sec	0.08	0.07	0.07	0.01	0.01	0.01
$\Delta = 1$ min	0.08	0.07	0.07	0.01	0.01	0.01
$\Delta = 5$ min	0.07	0.07	0.06	0.01	0.01	0.01
$\Delta = 30$ min	0.07	0.06	0.06	0.01	0.01	0.01
	$R = 500$					
True $Y_{t+1}^{\dagger}$	0.07	0.08	0.06	0.01	0.02	0.01
$\Delta = 5$ sec	0.08	0.08	0.06	0.01	0.02	0.01
$\Delta = 1$ min	0.07	0.08	0.06	0.01	0.02	0.01
$\Delta = 5$ min	0.07	0.08	0.06	0.01	0.02	0.01
$\Delta = 30$ min	0.06	0.07	0.05	0.01	0.02	0.01
	$R = 1000$					
True $Y_{t+1}^{\dagger}$	0.09	0.07	0.06	0.02	0.01	0.01
$\Delta = 5$ sec	0.09	0.07	0.06	0.02	0.01	0.01
$\Delta = 1$ min	0.09	0.07	0.06	0.02	0.01	0.01
$\Delta = 5$ min	0.08	0.07	0.06	0.03	0.01	0.01
$\Delta = 30$ min	0.07	0.06	0.05	0.02	0.01	0.01

Table 1: 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 and Vogelsang's "fixed-b" asymptotics.

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.[15]

	GW–NW			GW–KV		
Proxy BV_{t+1}^{Δ}	$P = 500$	$P = 1000$	$P = 2000$	$P = 500$	$P = 1000$	$P = 2000$
	$R = 250$					
True $Y_{t+1}^{\dagger}$	0.08	0.09	0.07	0.02	0.01	0.01
$\Delta = 5$ sec	0.08	0.09	0.07	0.02	0.01	0.01
$\Delta = 1$ min	0.08	0.09	0.06	0.02	0.01	0.01
$\Delta = 5$ min	0.07	0.07	0.06	0.02	0.01	0.01
$\Delta = 30$ min	0.04	0.04	0.04	0.01	0.01	0.01
	$R = 500$					
True $Y_{t+1}^{\dagger}$	0.09	0.08	0.07	0.01	0.01	0.01
$\Delta = 5$ sec	0.09	0.08	0.07	0.01	0.01	0.01
$\Delta = 1$ min	0.08	0.07	0.07	0.01	0.01	0.01
$\Delta = 5$ min	0.08	0.07	0.05	0.01	0.02	0.02
$\Delta = 30$ min	0.04	0.03	0.03	0.01	0.01	0.01
	$R = 1000$					
True $Y_{t+1}^{\dagger}$	0.09	0.08	0.07	0.01	0.01	0.01
$\Delta = 5$ sec	0.09	0.08	0.07	0.01	0.01	0.01
$\Delta = 1$ min	0.08	0.07	0.07	0.01	0.01	0.01
$\Delta = 5$ min	0.06	0.07	0.07	0.02	0.01	0.01
$\Delta = 30$ min	0.03	0.03	0.04	0.01	0.01	0.01

Table 2: 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 and Vogelsang's "fixed-b" asymptotics.

6.2 Results

The results for Simulations A, B and C are presented in Tables 1, 2 and 3, 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

[15] Due to the complexity from the data generating processes and volatility models we consider, computing the population MSE analytically for each forecast sequence is difficult. We instead compute the population MSE by simulation, using a Monte Carlo sample of 500,000 days. Similarly, it is difficult to construct data generating processes under which two forecast sequences have identical population MSE, which motivates our considering a nonzero χ in the null hypothesis, equation (2.2), of our simulation design. Doing so enables us to use realistic data generating processes and reasonably sophisticated forecasting models which mimic those used in prior empirical work.

	GW–NW			GW–KV		
Proxy RC_{t+1}^{Δ}	$P = 500$	$P = 1000$	$P = 2000$	$P = 500$	$P = 1000$	$P = 2000$
			$R = 250$			
True $Y_{t+1}^{\dagger}$	0.25	0.22	0.21	0.07	0.04	0.04
$\Delta = 5$ sec	0.25	0.22	0.21	0.07	0.04	0.04
$\Delta = 1$ min	0.25	0.23	0.20	0.07	0.04	0.04
$\Delta = 5$ min	0.24	0.23	0.20	0.06	0.05	0.04
$\Delta = 30$ min	0.24	0.21	0.19	0.07	0.05	0.04
			$R = 500$			
True $Y_{t+1}^{\dagger}$	0.29	0.27	0.24	0.12	0.06	0.05
$\Delta = 5$ sec	0.29	0.27	0.24	0.12	0.06	0.05
$\Delta = 1$ min	0.29	0.27	0.24	0.12	0.06	0.05
$\Delta = 5$ min	0.29	0.28	0.24	0.12	0.06	0.05
$\Delta = 30$ min	0.30	0.26	0.23	0.12	0.07	0.05
			$R = 1000$			
True $Y_{t+1}^{\dagger}$	0.27	0.23	0.20	0.14	0.07	0.06
$\Delta = 5$ sec	0.27	0.23	0.20	0.14	0.07	0.06
$\Delta = 1$ min	0.27	0.23	0.20	0.14	0.07	0.06
$\Delta = 5$ min	0.27	0.23	0.19	0.14	0.07	0.06
$\Delta = 30$ min	0.27	0.23	0.19	0.14	0.07	0.06

Table 3: 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 and Vogelsang's "fixed-b" asymptotics.

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 KV method is of interest here because of the well-known size distortion problem for inference procedures based on the standard HAC estimation theory; see Müller (2012) and references therein. We set the truncation lag to be $3P^{1/3}$ for NW and to be $0.5P$ for KV.[16]

Overall, we find that the rejection rates of the feasible tests based on proxies are generally

[16] In the KV case, the one-sided critical value for the t-statistic is 2.774 at 5% level when the truncation lag is $0.5P$; see Table 1 in Kiefer and Vogelsang (2005).

very close to the rejection rates of the infeasible tests using the true forecast target, and thus that our negligibility result holds well in a range of realistic simulation scenarios. The standard GW-NW method has reasonable size control in Simulations A and B, but has nontrivial size distortion for Simulation C.[17] This size distortion occurs even when the true target variable is used, and is not exacerbated by the use of proxies. The GW-KV method has better size control in these simulation scenarios, being somewhat conservative in Simulations A and B, and having good rejection rates in Simulation C for $P = 1000$ and $P = 2000$. Supplemental Appendix S.C presents results that confirm that these findings are robust with respect to the choice of the truncation lag in the estimation of the long-run variance, along with some additional results on the disagreement between the feasible and the infeasible tests.

It is perhaps surprising that our negligibility argument performs well in the simulations even when the sampling frequency for the realized measure is as low as 30 minutes, given that proxies formed using relatively low frequency returns are often imprecise. To see why this is the case, we emphasize that the negligibility result does *not* require the proxy be precise on every day of the sample (although that would of course be sufficient). Rather, it requires a much weaker condition (Assumption C), namely that the expectation of the proxy evaluation measure is precise in the evaluation sample. In our testing context, the precision of the expectation of the proxy evaluation measure is implicitly compared with the sampling variability of the sample mean of the evaluation function, $\bar{f}_T$. In applications commonly encountered in practice, the target variable is difficult to predict and persistent, forecast errors are occasionally quite large, and the length of evaluation samples are limited by the availability of high frequency data. All of these features lead to relatively large sampling variation in $\bar{f}_T$, and the simulation results in this section indicate that this sampling uncertainty dominates the difference in the expectation of the evaluation measure using the proxy rather than the true target variable.

[17] The reason for the large size distortion of the NW method in Simulation C appears to be the relatively high persistence in the quadratic loss differentials. 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.

7 Application: Comparing correlation forecasts

7.1 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., Patton (2011)) cannot be used either, due to the absence of any such proxy.[18] This is thus an ideal example to illustrate the usefulness of the method proposed in the current paper.

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.[19] 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 conditional mean and variance: for stock i, $i = 1$ or 2, its daily logarithmic return r_{it} follows

$$\begin{cases} r_{it} = \mu_i + \sigma_{it}\varepsilon_{it}, \\ \sigma_{it}^2 = \omega_i + \beta_i\sigma_{i,t-1}^2 + \alpha_i\sigma_{i,t-1}^2\varepsilon_{i,t-1}^2 + \delta_i\sigma_{i,t-1}^2\varepsilon_{i,t-1}^2 1_{\{\varepsilon_{i,t-1}\leq 0\}} + \gamma_i RV_{i,t-1}^{1\,\mathrm{min}}. \end{cases} \tag{7.1}$$

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 (6.11). The other three models are extensions of the baseline model. The first extension is

[18] 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.

[19] For all sampling intervals we use the "subsample-and-average" estimator of Zhang, Mykland, and Aït-Sahalia (2005b), with five subsamples when $\Delta = 5$ seconds, and with ten equally-spaced subsamples for the other choices of sampling frequency.

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:

$$Q_t = \overline{Q}\ (1 - a - b - d) + b\, Q_{t-1} + a\, \varepsilon_{t-1}\varepsilon_{t-1}^{\intercal} + d\, \eta_{t-1}\eta_{t-1}^{\intercal}, \quad \text{where} \quad \eta_t \equiv \varepsilon_t \circ 1_{\{\varepsilon_t \leq 0\}}. \tag{7.2}$$

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:

$$Q_t = \overline{Q}\ (1 - a - b - g) + b\, Q_{t-1} + a\, \varepsilon_{t-1}\varepsilon_{t-1}^{\intercal} + g\, RC_{t-1}^{65\,\mathrm{min}}. \tag{7.3}$$

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

$$Q_t = \overline{Q}\ (1 - a - b - d - g) + b\, Q_{t-1} + a\, \varepsilon_{t-1}\varepsilon_{t-1}^{\intercal} + d\, \eta_{t-1}\eta_{t-1}^{\intercal} + g\, RC_{t-1}^{65\,\mathrm{min}}. \tag{7.4}$$

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.

7.2 Results

Table 4 presents results for comparisons of each of the three generalized models and the baseline DCC model, using both the GW–NW and the GW–KV tests. The results in the first and fourth columns indicate that the A-DCC model does not improve predictive accuracy relative to the baseline DCC model. The GW–KV tests reveal that the loss of the A-DCC forecast is not statistically different from that of DCC. The GW–NW tests, on the other hand, report statistically significant outperformance of the A-DCC model relative to the DCC for some proxies, however this finding should 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, the more general A-DCC model actually underperforms the baseline model, though the difference is not significant. The next columns reveal that the R-DCC model outperforms the DCC model, particularly for the MSFT–AAPL pair, where the finding is highly significant and robust to the choice of proxy.

	GW–NW			GW–KV		
Proxy RC^{Δ}_{t+1}	DCC vs A-DCC	DCC vs R-DCC	DCC vs AR-DCC	DCC vs A-DCC	DCC vs R-DCC	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 4: 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.

Finally, we find that the AR-DCC model outperforms the DCC model, however 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 conservatively that the AR-DCC is not significantly better than the baseline DCC model.

Table 5 presents results from pairwise comparisons among the generalized models. Consistent with the results in Table 4, 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 yields mixed, but statistically insignificant, findings across the two pairs of stocks.

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.

	GW–NW			GW–KV		
Proxy RC_{t+1}^{Δ}	A-DCC vs R-DCC	A-DCC vs AR-DCC	R-DCC vs AR-DCC	A-DCC vs R-DCC	A-DCC vs AR-DCC	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 5: 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.

8 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, for example, Diebold and Mariano (1995), West (1996), White (2000), Giacomini and White (2006) and McCracken (2007), in cases where the latent target variable is replaced by a proxy computed using high-frequency (intraday) data. We derive convergence rate results for general classes of high-frequency based estimators of volatility and jump functionals, which cover a majority of 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. Based on these results, we provide conditions under which the moments that define the proxy hypotheses converge sufficiently quickly to their counterparts under the true hypotheses, so that the feasible tests based on proxies are valid under not only the former, but also the latter. 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.

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 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): "Testing for Jumps in a Discretely Observed Process," *Annals of Statistics*, 37, 184–222.

Aït-Sahalia, Y., P. A. Mykland, and L. Zhang (2005): "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise," *Review of Financial Studies*, 18, 351–416.

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.

Bandi, F., and R. Renò (2012): "Time-varying Leverage Effects," *Journal of Econometrics*, 169, 94–113.

Bandi, F. M., and J. R. Russell (2008): "Microstructure Noise, Realized Volatility and Optimal Sampling," *Review of Economic Studies*, 75, 339–369.

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.

Bollerslev, T., and H. Zhou (2002): “Estimating Stochastic Volatility Diffusions Using Conditional Moments of Integrated Volatility,” *Journal of Econometrics*, 109, 33–65.

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.

Comte, F., and E. Renault (1998): “Long Memory in Continuous-time Stochastic Volatility Models,” *Mathematical Finance*, 8, 291–323.

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

Duffie, D. (2001): *Dynamic Asset Pricing Theory*. Princeton University Press, third edn.

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.

Foster, D., and D. B. Nelson (1996): "Continuous record asymptotics for rolling sample variance estimators," *Econometrica*, 64, 139–174.

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

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. (2008): "Asymptotic properties of realized power variations and related functionals of semimartingales," *Stochastic Processes and their Applications*, 118, 517–559.

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

Jacod, J., and M. Rosenbaum (2013): "Quarticity and Other Functionals of Volatility: Efficient Estimation," *Annals of Statistics*, 118, 1462–1484.

Kanaya, S., and D. Kristensen (2010): "Estimation of Stochastic Volatility Models by Nonparametric Filtering," Discussion paper, University of Oxford.

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

Kristensen, D. (2010): "Nonparametric Filtering of the Realized Spot Volatility: A Kernel-Based Approach," *Econometric Theory*, 26(1), pp. 60–93.

Lepingle, D. (1976): "La Variation d'Ordre p des Semi-Martingales," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 36, 295–316.

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.

Müller, U. (2012): "HAC Corrections for Strongly Autocorrelated Time Series," Discussion paper, Princeton University.

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 K. Sheppard (2013): "Good volatility, bad volatility: Signed jumps and the persistence of volatility," *Review of Economics and Statistics*, Forthcoming.

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.

Renò, R. (2006): "Nonparametric estimation of stochastic volatility models," *Economics Letters*, 90, 390–395.

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

Singleton, K. J. (2006): *Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment*. Princeton University Press.

Todorov, V. (2009): "Estimation of Continuous-Time Stochastic Volatility Models with Jumps using High-Frequency Data," *Journal of Econometrics*, 148, 131–148.

Todorov, V., and G. Tauchen (2012): "The Realized Laplace Transform of Volatility," *Econometrica*, 80, 1105–1127.

Todorov, V., G. Tauchen, and I. Grynkiv (2011): "Realized Laplace Transforms for Estimation of Jump Diffusive Volatility Models," *Journal of Econometrics*, 164, 367–381.

van der Vaart, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press.

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. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

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

White, H. (2001): *Asymptotic Theory for Econometricians*. Academic Press.

Zhang, L. (2006): "Efficient Estimation of Stochastic Volatility Using Noisy Observations: A Multi-Scale Approach," *Bernoulli*, 12, 1019–1043.

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

——— (2005b): "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data," *Journal of the American Statistical Association*, 100, 1394–1411.