Comparing Predictive Accuracy under Long Memory - With an Application to Volatility Forecasting -*

Robinson Kruse[a,b] Christian Leschinski[c]
Michael Will[c]

[a]University of Cologne [b]CREATES, Aarhus University and [c]Leibniz University Hannover

October 30, 2017

Abstract

This paper extends the popular Diebold-Mariano test for equal predictive accuracy to situations when the forecast error loss differential exhibits long memory. This situation can arise frequently since long memory can be transmitted from forecasts and the forecast objective to forecast error loss differentials. The nature of this transmission depends on the (un)biasedness of the forecasts and whether the involved series share common long memory. Further theoretical results show that the conventional Diebold-Mariano test is invalidated under these circumstances. Robust statistics based on a memory and autocorrelation consistent estimator and an extended fixed-bandwidth approach are considered. The subsequent extensive Monte Carlo study provides numerical results on various issues. As empirical applications, we consider recent extensions of the HAR model for the S&P500 realized volatility. While we find that forecasts improve significantly if jumps are considered, improvements achieved by the inclusion of an implied volatility index turn out to be insignificant.

Key words: Equal Predictive Accuracy · Long Memory · Diebold-Mariano Test · Long-run Variance Estimation · Realized Volatility.

JEL classification: C22; C52; C53

*We would like to thank the editor Andrew Patton and two anonymous referees for their helpful remarks which improved the quality of the paper significantly. Moreovers, we thank Philipp Sibbertsen, Karim Abadir, Guillaume Chevillion, Mauro Costantini, Matei Demetrescu, Niels Haldrup, Uwe Hassler, Tucker McElroy and Uta Pigorsch as well as the participants of the 3rd Time Series Workshop in Rimini, the 2nd IAAE conference in Thessaloniki, the Statistische Woche 2015 in Hamburg, the 4th Long-Memory Symposium in Aarhus, the 16th IWH-CIREQ Workshop in Halle and the CFE 2015 in London for their helpful comments. Robinson Kruse gratefully acknowledge support from CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation.

1 Introduction

If the accuracy of competing forecasts is to be evaluated in a (pseudo-)out-of-sample setup, it has become standard practice to employ the test of Diebold and Mariano (1995) (hereafter DM test). Let $\widehat{y}_{1t}$ and $\widehat{y}_{2t}$ denote two competing forecasts for the forecast objective series y_t and let the loss function be given by $g(y_t, \widehat{y}_{it}) \geq 0$ for $i = 1, 2$. The forecast error loss differential is then denoted by

$$z_t = g(y_t, \widehat{y}_{1t}) - g(y_t, \widehat{y}_{2t}). \tag{1}$$

By only imposing restrictions on the loss differential z_t, instead of the forecast objective and the forecasts, Diebold and Mariano (1995) test the null hypothesis of equal predictive accuracy, i.e. $H_0 : E(z_t) = 0$, by means of a simple t-statistic for the mean of the loss differentials. In order to account for serial correlation, a long-run variance estimator such as the heteroscedasticity and autocorrelation consistent (HAC) estimator is applied (see Newey and West (1987), Andrews (1991) and Andrews and Monahan (1992)). For weakly dependent and second-order stationary processes this leads to an asymptotic standard normal distribution of the t-statistic.

Apart from the development of other forecast comparison tests such as those of West (1996) or Giacomini and White (2006), several direct extensions and improvements of the DM test have been proposed. Harvey et al. (1997) suggest a version that corrects for the bias of the long-run variance estimation in finite samples. A multivariate DM test is derived by Mariano and Preve (2012). To mitigate the well known size issues of HAC-based tests in finite samples of persistent short memory processes, Choi and Kiefer (2010) construct a DM test using the so-called fixed-bandwidth (or in short, fixed-b) asymptotics, originally introduced in Kiefer and Vogelsang (2005) (see also Li and Patton (2015)). The issue of near unit root asymptotics is tackled by Rossi (2005). These studies belong to the classical $I(0)/I(1)$ framework.

Contrary to the aforementioned studies, we consider the situation in which the loss differentials follow long memory processes. Our first contribution is to show that long memory can be transmitted from the forecasts and the forecast objective to the forecast errors and subsequently to the forecast error loss differentials. We provide theoretical results for the mean squared error (MSE) loss function and Gaussian processes. We give conditions under which the transmission occurs and characterize the memory properties of the forecast error loss differential. The memory transmission for non-Gaussian processes and other loss functions is demonstrated by means of Monte Carlo simulations resembling typical forecast scenarios. As a second contribution, we show (both theoretically and via simulations) that the original DM test is invalidated under long memory and suffers from severe upward size distortions.

Third, we study two simple extensions of the DM statistic that permit valid inference under long and short memory. These extensions are based on the memory and autocorrelation consistent (MAC) estimator of Robinson (2005) (see also Abadir et al. (2009)) and the extended fixed-b asymptotics (EFB) of McElroy and Politis (2012). The performance of these modified statistics is analyzed in a Monte Carlo study that is specifically tailored to reflect the properties that are likely to occur in the loss differentials. We compare several bandwidth and kernel choices that allow recommendations for practical applications.

Our fourth contribution is an empirical application in which we reconsider two recent extensions of the heterogeneous autoregressive model for realized volatility (HAR-RV) by Corsi (2009). First, we test whether forecasts obtained from HAR-RV type models can be improved by including information on model-free risk-neutral implied volatility which is measured by the CBOE volatility index (VIX). We find that short memory approaches (classic Diebold-Mariano test and fixed-b versions) reject the null hypothesis of equal predictive accuracy in favor of models including implied volatility. On the contrary, our long memory robust statistics do not indicate a significant improvement in forecast performance which implies that previous rejections might be spurious due to neglected long memory.

The second issue we tackle in our empirical applications relates to earlier work by inter alia Andersen et al. (2007) and Corsi et al. (2010), who consider the decomposition of the quadratic variation of the log-price process into a continuous integrated volatility component and a discrete jump component. Here, we find that the separate treatment of continuous components and jump components significantly improves forecasts of realized variance for short forecast horizons even if the memory in the loss differentials is accounted for.

The rest of this paper is organized as follows. Section 2 reviews the classic Diebold-Mariano test and presents the fixed-b approach for the short memory case. Section 3 covers the case of long-range dependence and contains our theoretical results on the transmission of long memory to the loss differential series. Two distinct approaches to design a robust t-statistic are discussed in Section 4. Section 5 contains our Monte Carlo study and in Section 6 we present our empirical results. Conclusions are drawn in Section 7. All proofs are contained in the Appendix.

2 Diebold-Mariano Test

Diebold and Mariano (1995) construct a test for $H_0 : E\left[g(y_t, \widehat{y}_{1t}) - g(y_t, \widehat{y}_{2t})\right] = E(z_t) = 0$, solely based on assumptions on the loss differential series z_t. Suppose that z_t follows the weakly stationary linear process

$$z_t = \mu_z + \sum_{j=0}^{\infty} \theta_j v_{t-j} \ , \tag{2}$$

where it is required that $|\mu_z| < \infty$ and $\sum_{j=0}^{\infty} \theta_j^2 < \infty$ hold. For simplicity of the exposition we additionally assume that $v_t \sim iid(0, \sigma_v^2)$. If $\widehat{y}_{1t}$ and $\widehat{y}_{2t}$ are performing equally good in terms of $g(\cdot)$, $\mu_z = 0$ holds, otherwise $\mu_z \neq 0$. The corresponding t-statistic is based on the sample mean $\bar{z} = T^{-1} \sum_{t=1}^{T} z_t$ and an estimate $(\widehat{V})$ of the long-run variance $V = \lim_{T\to\infty} \mathrm{Var}\big(T^{\delta}\left(\bar{z} - \mu_z\right)\big)$. The DM statistic is given by

$$t_{DM} = T^{\delta} \frac{\bar{z}}{\sqrt{\widehat{V}}} \ . \tag{3}$$

Under stationary short memory, we have $\delta = 1/2$, while the rate changes to $\delta = 1/2 - d$ under stationary long memory, with $0 < d < 1/2$ being the long memory parameter. The (asymptotic) distribution of this t-statistic hinges on the autocorrelation properties of the loss differential

series z_t. In the following, we shall distinguish two cases: (1) z_t is a stationary short memory process and (2) strong dependence in form of a long memory process is present in z_t as presented in Section 3.

2.1 Conventional Approach: HAC

For the estimation of the long-run variance V, Diebold and Mariano (1995) suggest to use the truncated long-run variance of an MA($h-1$) process for an h-step-ahead forecast. This is motivated by the fact that optimal h-step-ahead forecast errors of a linear time series process follow an MA($h-1$) process. Nevertheless, as pointed out by Diebold (2015), among others, the test is readily extendable to more general situations if, for example, HAC estimators are used (see also Clark (1999) for some early simulation evidence). The latter have become the standard class of estimators for the long-run variance. In particular,

$$\widehat{V}_{HAC} = \sum_{j=-T+1}^{T-1} k\left(\frac{j}{B}\right) \widehat{\gamma}_z(j) \,, \tag{4}$$

where $k(\cdot)$ is a user-chosen kernel function, B denotes the bandwidth and

$$\widehat{\gamma}_z(j) = \frac{1}{T} \sum_{t=|j|+1}^{T} (z_t - \bar{z}) \left(z_{t-|j|} - \bar{z}\right)$$

is the usual estimator for the autocovariance of process z_t at lag j. The corresponding Diebold-Mariano statistic is given by

$$t_{HAC} = T^{1/2} \frac{\bar{z}}{\sqrt{\widehat{V}_{HAC}}}. \tag{5}$$

If z_t is weakly stationary with absolutely summable autocovariances $\gamma_z(j)$, it holds that $V = \sum_{j=-\infty}^{\infty} \gamma_z(j)$. Suppose that a central limit theorem applies for partial sums of z_t, so that $\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} z_t \Rightarrow \sqrt{V} W(r)$ where $W(r)$ is a standard Brownian motion. Then, the t_{HAC}-statistic is asymptotically standard normal under the null hypothesis, i.e.

$$t_{HAC} \Rightarrow \mathcal{N}(0,1).$$

For the sake of a comparable notation to the long memory case, note that $V = 2\pi f_z(0)$, where $f_z(0)$ is the spectral density function of z_t at frequency zero.

2.2 Fixed-bandwidth Approach

Even though nowadays the application of HAC estimators is standard practice, related tests are often found to be seriously size-distorted in finite samples, especially under strong persistence. It is assumed that the ratio $b = B/T \to 0$ as $T \to \infty$ in order to achieve a consistent estimation of the long-run variance V (see for instance Andrews (1991) for additional technical details). Kiefer and Vogelsang (2005) develop a new asymptotic framework in which the ratio B/T approaches a fixed constant $b \in (0,1]$ as $T \to \infty$. Therefore, it is called fixed-b inference as opposed to the

classical small-b HAC approach where $b \to 0$.

In the case of fixed-b (FB), the estimator $\widehat{V}(k,b)$ does not converge to V any longer. Instead, $\widehat{V}(k,b)$ converges to V multiplied by a functional of a Brownian bridge process. In particular, $\widehat{V}(k,b) \Rightarrow VQ(k,b)$. Therefore, the corresponding t-statistic

$$t_{FB} = T^{1/2} \frac{\bar{z}}{\sqrt{\widehat{V}(k,b)}} \tag{6}$$

has a non-normal and non-standard limiting distribution, i.e.

$$t_{FB} \Rightarrow \frac{W(1)}{\sqrt{Q(k,b)}} \, .$$

Here, $W(r)$ is a standard Brownian motion on $r \in [0,1]$. Both, the choice of the bandwidth parameter b and the (twice continuously differentiable) kernel k appear in the limit distribution. For example, for the *Bartlett* kernel we have

$$Q(k,b) = \frac{2}{b} \left(\int_0^1 \widetilde{W}(r)^2 \mathrm{d}r - \int_0^{1-b} \widetilde{W}(r+b)\widetilde{W}(r) \mathrm{d}r \right) \, ,$$

with $\widetilde{W}(r) = W(r) - rW(1)$ denoting a standard Brownian bridge. Thus, critical values reflect the user choices on the kernel and the bandwidth even in the limit. In many settings, fixed-b inference is more accurate than the conventional HAC estimation approach. An example of its application to forecast comparisons are the aforementioned articles of Choi and Kiefer (2010) and Li and Patton (2015), who apply both techniques (HAC and fixed-b) to compare exchange rate forecasts. Our Monte Carlo simulation study sheds additional light on their relative empirical performance.

3 Long Memory in Forecast Error Loss Differentials

3.1 Preliminaries

Under long-range dependence in z_t, one has to expect that neither conventional HAC estimators nor the fixed-b approach can be applied without any further modification since strong dependence such as fractional integration is ruled out by assumption of a weakly stationary linear process. In particular, we show that HAC-based tests reject with probability one in the limit (as $T \to \infty$) if z_t has long memory. This results is stated in our Proposition 6 (at the end of this section). As our finite-sample simulations clearly demonstrate, this implies strong upward size distortions and invalidates the use of the classic DM test statistic. Before we actually state these results formally, we first show that the loss differential z_t may exhibit long memory in various situations. We start with a basic definition of stationary long memory time series.

Definition 1. *A time series a_t with spectral density $f_a(\lambda)$, for $\lambda \in [-\pi, \pi]$, has long memory with memory parameter $d_a \in (0, 1/2)$, if $f_a(\lambda) \sim L_f |\lambda|^{-2d_a}$, for $d_a \in (0, 1/2)$, as $\lambda \to 0$. The function $L_f(\cdot)$ is slowly varying at the origin. We then write $a_t \sim LM(d_a)$.*

This is the usual definition of a stationary long memory process and Theorem 1.3 of Beran et al. (2013) states that under this restriction and mild regularity conditions, Definition 1 is equivalent to $\gamma_a(j) \sim L_\gamma |j|^{2d_a-1}$ as $j \to \infty$, where $\gamma_a(j)$ is the autocovariance function of a_t at lag j and $L_\gamma(\cdot)$ is slowly varying at infinity. If $d_a = 0$ holds, the process has short memory. Our results build on the asymptotic behavior of the autocovariances that have the long memory property from Definition 1. Whether this memory is generated by fractional integration can not be inferred. However, this does not affect the validity of the test statistics introduced in Section 4. We therefore adopt Definition 1 which covers fractional integration as a special case. A similar approach is taken by Dittmann and Granger (2002).[1]

Given Definition 1, we now state some assumptions regarding the long memory structure of the forecast objective and the forecasts.

Assumption 1 (Long Memory)**.** *The time series y_t, $\widehat{y}_{1t}$, $\widehat{y}_{2t}$ with expectations $E(y_t) = \mu_y$, $E(\widehat{y}_{1t}) = \mu_1$ and $E(\widehat{y}_{2t}) = \mu_2$ are causal Gaussian long memory processes (according to Definition 1) of orders d_y, d_1 and d_2, respectively.*

Similar to Dittmann and Granger (2002), we rely on the assumption of Gaussianity since no results for the memory structure of squares and cross-products of non-Gaussian long memory processes are available in the existing literature. It shall be noted that Gaussianity is only assumed for the derivation of the memory transmission from the forecasts and the forecast objective to the loss differential, but not for the subsequent results.

In the following, we make use of the concept of common long memory in which a linear combination of long memory series has reduced memory. The amount of reduction is labeled as b in accordance with the literature (similar to the symbol b in "fixed-b", but no confusion shall arise).

Definition 2 (Common Long Memory)**.** *The time series a_t and b_t have common long memory (CLM) if both a_t and b_t are $LM(d)$ and there exists a linear combination $c_t = a_t - \psi_0 - \psi_1 b_t$ with $\psi_0 \in \mathbb{R}$ and $\psi_1 \in \mathbb{R}\backslash 0$ such that $c_t \sim LM(d-b)$, for some $d \geq b > 0$. We write $a_t, b_t \sim CLM(d, d-b)$.*

For simplicity and ease of exposition, we first exclude the possibility of common long memory among the series. This assumption is relaxed later on.

Assumption 2 (No Common Long Memory)**.** *If $a_t, b_t \sim LM(d)$, then $a_t - \psi_0 - \psi_1 b_t \sim LM(d)$ for all $\psi_0 \in \mathbb{R}, \psi_1 \in \mathbb{R}$ and $a_t, b_t \in \{y_t, \widehat{y}_{1t}, \widehat{y}_{2t}\}$.*

In order to derive the long memory properties of the forecast error loss differential, we make use of a result in Leschinski (2017) that characterizes the memory structure of the product series $a_t b_t$ for two long memory time series a_t and b_t. Such products play an important role in the following analysis. The result is therefore shown as Proposition 1 below, for convenience.

[1] Sometimes the terms long memory and fractional integration are used interchangeably. However, a stationary fractionally integrated process a_t has spectral density $f_a(\lambda) = |1 - e^{i\lambda}|^{-2d_a} G_a(\lambda)$, so that $f_a(\lambda) \sim G(\lambda)|\lambda|^{-2d_a}$ as $\lambda \to 0$ since $|1 - e^{i\lambda}| \to \lambda$ as $\lambda \to 0$. Therefore, fractional integration is a special case of long memory, but many other processes would satisfy Definition 1, too. Examples include non-causal processes and processes with trigonometric power law coefficients, as recently discussed in Kechagias and Pipiras (2015).

Proposition 1 (Memory of Products)**.** *Let a_t and b_t be long memory series according to Definition 1 with memory parameters d_a and d_b, and means μ_a and μ_b, respectively. Then*

$$a_t b_t \sim \begin{cases} LM(\max\{d_a, d_b\}), & \text{for } \mu_a, \mu_b \neq 0 \\ LM(d_a), & \text{for } \mu_a = 0, \mu_b \neq 0 \\ LM(d_b), & \text{for } \mu_b = 0, \mu_a \neq 0 \\ LM(\max\{d_a + d_b - 1/2, 0\}), & \text{for } \mu_a = \mu_b = 0 \text{ and } S_{a,b} \neq 0 \\ LM(d_a + d_b - 1/2), & \text{for } \mu_a = \mu_b = 0 \text{ and } S_{a,b} = 0, \end{cases}$$

where $S_{a,b} = \sum_{j=-\infty}^{\infty} \gamma_a(j)\gamma_b(j)$ with $\gamma_a(\cdot)$ and $\gamma_b(\cdot)$ denoting the autocovariance functions of a_t and b_t, respectively.

Proposition 1 shows that the memory of products of long memory time series critically depends on the means μ_a and μ_b of the series a_t and b_t. If both series are mean zero, the memory of the product is either the maximum of the sum of the memory parameters of both factor series minus one half - or it is zero - depending on the sum of autocovariances. Since $d_a, d_b < 1/2$, this is always smaller than any of the original memory parameters. If only one of the series is mean zero, the memory of the product $a_t b_t$ is determined by the memory of this particular series. Finally, if both series have non-zero means, the memory of the product is equal to the maximum of the memory orders of the two series.

Furthermore, Proposition 1 makes a distinction between antipersistent series and short memory series if the processes have zero means and $d_a + d_b - 1/2 < 0$. Our results below, however, do not require this distinction. The reason being that a linear combination involving the square of at least one of the series appears in each case, and these cannot be anti-persistent long memory processes (see the proofs of Propositions 2 and 5 for details).

As discussed in Leschinski (2017), Proposition 1 is related to the results in Dittmann and Granger (2002), who consider the memory of non-linear transformations of zero mean long memory time series that can be represented through a finite sum of Hermite polynomials. Their results include the square a_t^2 of a time series which is also covered by Proposition 1 if $a_t = b_t$. If the mean is zero ($\mu_a = 0$), we have $a_t^2 \sim LM(\max\{2d_a - 1/2, 0\})$. Therefore, the memory is reduced to zero if $d \leq 1/4$. However, as can be seen from Proposition 1, this behavior depends critically on the expectation of the series.

Since it is the most widely used loss function in practice, we focus on the MSE loss function $g(y_t, \widehat{y}_{it}) = (y_t - \widehat{y}_{it})^2$ for $i = 1, 2$. The quadratic forecast error loss differential is then given by

$$z_t = (y_t - \widehat{y}_{1t})^2 - (y_t - \widehat{y}_{2t})^2 = \widehat{y}_{1t}^2 - \widehat{y}_{2t}^2 - 2y_t(\widehat{y}_{1t} - \widehat{y}_{2t}). \tag{7}$$

Even though the forecast objective y_t as well as the forecasts $\widehat{y}_{it}$ in (7) carry a time index t, the representation is quite versatile. It permits forecasts to be generated from time series models where $\widehat{y}_{it} = \sum_{s=1}^{t-1} \phi_s y_{t-s}$ as well as predictive regressions with $\widehat{y}_{it} = \beta' x_{t-s}$, where β is a parameter vector and x_{t-s} is a vector of explanatory variables lagged by s periods. In addition to that, even though estimation errors are not considered explicitly, they would be reflected by the fact that $E[y_t|\Psi_{t-h}] \neq \widehat{y}_{it|t-h}$, where Ψ_{t-h} is the information set available at the forecast

origin $t-h$. This means that forecasts are biased in presence of estimation error, even if the model employed corresponds to the true data generating process. The forecasts are also not restricted to be obtained from a linear model. Similar to the Diebold-Mariano test, which is solely based on a single assumption on the forecast error loss differential (7), the following results are derived by assuming certain properties of the forecasts and the forecast objective. Therefore, we follow Diebold and Mariano (1995) and do not impose direct restrictions on the way forecasts are generated.

3.2 Transmission of Long Memory to the Loss Differential

Following the introduction of the necessary definitions and a preliminary result, we now present the result for the memory order of z_t defined via (7) in Proposition 2. It is based on the memory of y_t, $\widehat{y}_{1t}$ and $\widehat{y}_{2t}$ and assumes the absence of common long memory for simplicity.

Proposition 2 (Memory Transmission without CLM)**.** *Under Assumptions 1 and 2, the forecast error loss differential in (7) is $z_t \sim LM(d_z)$, where*

$$d_z = \begin{cases} \max\{d_y,\ d_1,\ d_2\}, & \text{if } \mu_1 \neq \mu_2 \neq \mu_y \\ \max\{d_1,\ d_2\}, & \text{if } \mu_1 = \mu_2 \neq \mu_y \\ \max\{2d_1 - 1/2,\ d_2,\ d_y\}, & \text{if } \mu_1 = \mu_y \neq \mu_2 \\ \max\{2d_2 - 1/2,\ d_1,\ d_y\}, & \text{if } \mu_1 \neq \mu_y = \mu_2 \\ \max\big\{2\max\{d_1,\ d_2\} - 1/2,\ d_y + \max\{d_1,\ d_2\} - 1/2,\ 0\big\}, & \text{if } \mu_1 = \mu_2 = \mu_y. \end{cases}$$

Proof: *See the Appendix.*

The basic idea of the proof relates to Proposition 3 of Chambers (1998). It shows that the behavior of a linear combination of long memory series is dominated by the series with the strongest memory. Since we know from Proposition 1 that μ_1, μ_2 and μ_y play an important role for the memory of a squared long memory series, we set $y_t = y_t^* + \mu_y$ and $\widehat{y}_{it} = \widehat{y}_{it}^* + \mu_i$, so that the starred series denote the demeaned series and μ_i denotes the expected value of the respective series. Straightforward algebra yields

$$z_t = \widehat{y}_{1t}^{*2} - \widehat{y}_{2t}^{*2} - 2\Big[y_t^*(\mu_1 - \mu_2) + \widehat{y}_{1t}^*(\mu_y - \mu_1) + \widehat{y}_{2t}^*(\mu_y - \mu_2)\Big] - 2\Big[y_t^*(\widehat{y}_{1t}^* - \widehat{y}_{2t}^*)\Big] + const. \quad (8)$$

From (8), it is apparent that z_t is a linear combination of (i) the squared forecasts $\widehat{y}_{1t}^{*2}$ and $\widehat{y}_{2t}^{*2}$, (ii) the forecast objective y_t, (iii) the forecast series $\widehat{y}_{1t}^*$ and $\widehat{y}_{2t}^*$ and (iv) products of the forecast objective with the forecasts, i.e. $y_t^*\widehat{y}_{1t}^*$ and $y_t^*\widehat{y}_{2t}^*$. The memory of the squared series and the product series is determined in Proposition 1, from which the zero mean product series $y_t^*\widehat{y}_{it}^*$ is $LM(\max\{d_y + d_i - 1/2,\ 0\})$ or $LM(d_y + d_i - 1/2)$. Moreover, the memory of the squared zero mean series $\widehat{y}_{it}^{*2}$ is $\max\{2d_i - 1/2,\ 0\}$. By combining these results with that of Chambers (1998), the memory of the loss differential z_t is the maximum of all memory parameters of the components in (8). Proposition 2 then follows from a case-by-case analysis.

It demonstrates the transmission of long memory from the forecasts $\widehat{y}_{1t}$, $\widehat{y}_{2t}$ and the forecast objective y_t to the loss differential z_t. The nature of this transmission, however, critically hinges on the (un)biasedness of the forecasts. If both forecasts are unbiased (i.e. if $\mu_1 = \mu_2 = \mu_y$), the memory from all three input series is reduced and the memory of the loss differential z_t is equal to the maximum of (i) these reduced orders and (ii) zero. Therefore, only if memory parameters are small enough such that $d_y + \max\{d_1 + d_2\} < 1/2$, the memory of the loss differential z_t is reduced to zero. In all other cases, there is a transmission of dependence from the forecast and/or the forecast objective to the loss differential. The reason for this can immediately be seen from (8). Terms in the first bracket have larger memory than the remaining ones, because $d_i > 2d_i - 1/2$ and $\max\{d_y, d_i\} > d_y + d_i - 1/2$. Therefore, these terms dominate the memory of the products and squares whenever biasedness is present, i.e. $\mu_i - \mu_y \neq 0$ holds. Interestingly, the transmission of memory from the forecast objective y_t is prevented if both forecasts have equal bias, i.e. $\mu_1 = \mu_2$. On the contrary, if $\mu_1 \neq \mu_2$, d_z is at least as high as d_y.

3.3 Memory Transmission under Common Long Memory

The results in Proposition 2 are based on Assumption 2 that precludes common long memory among the series. Of course, in practice it is likely that such an assumption is violated. In fact, it can be argued that reasonable forecasts of long memory time series should have common long memory with the forecast objective. Therefore, we relax this restrictive assumption and replace it with Assumption 3, below.

Assumption 3 (Common Long Memory)**.** *The causal Gaussian process x_t has long memory according to Definition 1 of order d_x with expectation $E(x_t) = \mu_x$. If $a_t, b_t \sim CLM(d_x, d_x - b)$, then they can be represented as $y_t = \beta_y + \xi_y x_t + \eta_t$ for $a_t, b_t = y_t$ and $\widehat{y}_{it} = \beta_i + \xi_i x_t + \varepsilon_{it}$, for $a_t, b_t = \widehat{y}_{it}$, with $\xi_y, \xi_i \neq 0$. Both, η_t and ε_{it} are mean zero causal Gaussian long memory processes with parameters d_η and d_{ε_i} fulfilling $1/2 > d_x > d_\eta, d_{\varepsilon_i} \geq 0$, for $i = 1, 2$.*

Assumption 3 restricts the common long memory to be of a form so that both series a_t and b_t can be represented as linear functions of their joint factor x_t. This excludes more complicated forms of dependence that are sometimes considered in the cointegration literature such as non-linear or time-varying cointegration.

We know from Proposition 2 that the transmission of memory critically depends on the biasedness of the forecasts which leads to a complicated case-by-case analysis. If common long memory according to Assumption 3 is allowed for, we have an even more complex situation since there are several possible relationships: CLM of y_t with one of the $\widehat{y}_{it}$; CLM of both $\widehat{y}_{it}$ with each other, but not with y_t; and CLM of each $\widehat{y}_{it}$ with y_t. Each of these situations has to be considered with all possible combinations of the ξ_a and the μ_a for all $a \in \{y, 1, 2\}$.

To deal with this complexity, we focus on three important special cases: (i) the forecasts are biased and the ξ_a differ from each other, (ii) the forecasts are biased, but the ξ_a are equal, and (iii) the forecasts are unbiased and $\xi_a = \xi_b$ if a_t and b_t are in a common long memory relationship.

To understand the role of the coefficients ξ_a and ξ_b in the series that are subject to CLM, note that the forecast errors $y_t - \widehat{y}_{it}$ impose a cointegrating vector of $(1, -1)$. A different scaling of

the forecast objective and the forecasts is not possible. In the case of CLM between y_t and $\widehat{y}_{it}$, for example, we have from Assumption 3 that

$$y_t - \widehat{y}_{it} = \beta_y - \beta_i + x_t(\xi_y - \xi_i) + \eta_t - \varepsilon_{it}, \tag{9}$$

so that $x_t(\xi_y - \xi_i)$ does not disappear from the linear combination if the scaling parameters ξ_y and ξ_i are different from each other. We refer to a situation where $\xi_a = \xi_b$ as "balanced CLM", whereas CLM with $\xi_a \neq \xi_b$ is referred to as "unbalanced CLM".

In the special case (i) both forecasts are biased and the presence of CLM does not lead to a cancellation of the memory of x_t in the loss differential. Of course this can be seen as an extreme case, but it serves to illuminate the mechanisms at work - especially in contrast to the results in Propositions 4 and 5, below. By substituting the linear relations from Assumption 3 for those series involved in the CLM relationship in the loss differential $z_t = \widehat{y}_{1t}^2 - \widehat{y}_{2t}^2 - 2y_t(\widehat{y}_{1t} - \widehat{y}_{2t})$ and again setting $a_t = a_t^* + \mu_a$ for those series that are not involved in the CLM relationship, it is possible to find expressions that are analogous to (8). Since analogous terms to those in the first bracket of (8) appear in each case, it is possible to focus on the transmission of memory from the forecasts and the objective function to the loss differential. We obtain the following result.

Proposition 3 (Memory Transmission with Biased Forecasts and Unbalanced CLM)**.** *Let $\xi_i \neq \xi_y$, $\xi_1 \neq \xi_2$, $\mu_i \neq \mu_y$, and $\mu_1 \neq \mu_2$, for $i = 1, 2$. Then under Assumptions 1 and 3, the forecast error loss differential in (7) is $z_t \sim LM(d_z)$, where*

$$d_z = \begin{cases} \max\{d_y, d_x\}, & \text{if } \widehat{y}_{1t}, \widehat{y}_{2t} \sim CLM(d_x, d_x - b), \text{ except if } \xi_1/\xi_2 = (\mu_y - \mu_2)/(\mu_y - \mu_1) \\ \max\{d_2, d_x\}, & \text{if } \widehat{y}_{1t}, y_t \sim CLM(d_x, d_x - b), \text{ except if } \xi_1/\xi_y = -(\mu_1 - \mu_2)/(\mu_y - \mu_1) \\ \max\{d_1, d_x\}, & \text{if } \widehat{y}_{2t}, y_t \sim CLM(d_x, d_x - b), \text{ except if } \xi_2/\xi_y = -(\mu_1 - \mu_2)/(\mu_y - \mu_2) \\ d_x, & \text{if } \widehat{y}_{1t}, \widehat{y}_{2t}, y_t \sim CLM(d_x, d_x - b), \\ & \text{except if } \xi_1(\mu_y - \mu_1) + \xi_y(\mu_1 - \mu_2) = \xi_2(\mu_y - \mu_2). \end{cases}$$

Proof: *See the Appendix.*

In absence of fractional cointegration we observed in Proposition 1 that the memory is given as $\max\{d_1, d_2, d_y\}$ if the means differ from each other. Now, if two of the series are fractionally cointegrated, they both have memory d_x. Hence, Proposition 3 shows that the transmission mechanism is essentially unchanged and the memory of the loss differential is still dominated by the largest memory parameter. The only exception to this rule is if - by coincidence - the differences in the means and the memory parameters offset each other.

Similar to (i), case (ii) refers to a situation of biasedness, but now with balanced CLM, so that the underlying long memory factor x_t cancels out in the forecast error loss differentials. The memory transmission can than be characterized by the following proposition.

Proposition 4 (Memory Transmission with Biased Forecasts and Balanced CLM)**.** *Let $\xi_1 = \xi_2 = \xi_y$. Then under Assumptions 1 and 3, the forecast error loss differential in (7) is $z_t \sim$*

$LM(d_z)$, where

$$d_z = \begin{cases} \max\{d_y, d_x\}, & \text{if } \widehat{y}_{1t}, \widehat{y}_{2t} \sim CLM(d_x, d_x - b), \text{ and } \mu_1 \neq \mu_2 \\ \max\{d_2, d_x\}, & \text{if } \widehat{y}_{1t}, y_t \sim CLM(d_x, d_x - b), \text{ and } \mu_y \neq \mu_2 \\ \max\{d_1, d_x\}, & \text{if } \widehat{y}_{2t}, y_t \sim CLM(d_x, d_x - b), \text{ and } \mu_y \neq \mu_1 \\ \tilde{d}, & \text{if } \widehat{y}_{1t}, \widehat{y}_{2t}, y_t \sim CLM(d_x, d_x - b), \end{cases}$$

for some $0 \leq \tilde{d} < d_x$.

Proof: *See the Appendix.*

We refer to the first three cases in Propositions 3 and 4 as "partial CLM" as there is always one of the $\widehat{y}_{it}$ or y_t that is not part of the CLM relationship and the fourth case as "full CLM". We can observe that the dominance of the memory of the most persistent series under partial CLM is preserved for both balanced and unbalanced CLM. We therefore conclude that this effect is generated by the interaction with the series that is not involved in the CLM relationship. This can also be seen from equations (22) to (24) in the proof.

Only in the fourth case with full CLM, the memory transmission changes between Propositions 3 and 4. In this case, the memory in the loss differential is reduced to $d_z < d_x$.

The third special case (iii) refers to a situation of unbiasedness similar to the last case in Proposition 2. In addition to that, it is assumed that there is balanced CLM as in Proposition 4, where $\xi_a = \xi_b$ if a_t and b_t are in a common long memory relationship. Compared to the setting of the previous propositions this is the most ideal situation in terms of forecast accuracy. Here, we have the following result.

Proposition 5 (Memory Transmission with Unbiased Forecasts and Balanced CLM)**.** *Under Assumptions 1 and 3, and if $\mu_y = \mu_1 = \mu_2$ and $\xi_y = \xi_a = \xi_b$, then $z_t \sim LM(d_z)$, with*

$$d_z = \begin{cases} \max\{d_2 + \max\{d_x, d_\eta\} - 1/2,\ 2\max\{d_x, d_2\} - 1/2,\ d_{\varepsilon_1}\}, & \text{if } y_t, \widehat{y}_{1t} \sim CLM(d_x, d_x - \widetilde{b}) \\ \max\{d_1 + \max\{d_x, d_\eta\} - 1/2,\ 2\max\{d_x, d_1\} - 1/2,\ d_{\varepsilon_2}\}, & \text{if } y_t, \widehat{y}_{2t} \sim CLM(d_x, d_x - \widetilde{b}) \\ \max\{\max\{d_x,\ d_y\} + \max\{d_{\varepsilon_1},\ d_{\varepsilon_2}\} - 1/2,\ 0\}, & \text{if } \widehat{y}_{1t}, \widehat{y}_{2t} \sim CLM(d_x, d_x - \widetilde{b}) \\ \max\{d_\eta + \max\{d_{\varepsilon_1},\ d_{\varepsilon_2}\} - 1/2,\ 2\max\{d_{\varepsilon_1},\ d_{\varepsilon_2}\} - 1/2,\ 0\}, & \text{if } y_t, \widehat{y}_{1t} \sim CLM(d_x, d_x - \widetilde{b}) \\ & \text{and } y_t, \widehat{y}_{2t} \sim CLM(d_x, d_x - \widetilde{b}). \end{cases}$$

Here, $0 < \widetilde{b} \leq 1/2$ denotes a generic constant for the reduction in memory.

Proof: *See the Appendix.*

Proposition 5 shows that the memory of the forecasts and the objective variable can indeed cancel out if the forecasts are unbiased and if they have the same factor loading on x_t (i.e. if $\xi_1 = \xi_2 = \xi_y$). However, in the first two cases, the memory of the error series ε_{1t} and ε_{2t} imposes a lower bound on the memory of the loss differential. Furthermore, even though the memory can be reduced to zero in the third and fourth case, this situation only occurs if the memory

orders of x_t, y_t and the error series are sufficiently small. Otherwise, the memory is reduced, but does not vanish.

Overall, the results in Propositions 2, 3, 4 and 5 show that long memory can be transmitted from forecasts or the forecast objective to the forecast error loss differentials. Our results also show that the biasedness of the forecasts plays an important role for the transmission of dependence to the loss differentials.

To get further insights into the mechanisms found in Propositions 2, 3, 4 and 5 let us consider a situation in which two forecasts with different non-zero biases are compared. In absence of CLM, it is obvious from Proposition 2 that the memory of the loss differential is determined by the maximum of the memory orders of the forecasts and the forecast objective. If one of the forecasts has common long memory with the objective, the same holds true - irrespective of the loadings ξ_a on the common factor. As can be seen from Proposition 3, even if both forecasts have CLM with the objective, the maximal memory order is transmitted to z_t if the factor loadings ξ_a differ. Only if the factor loadings are equal, the memory is reduced as stated Proposition 4. If we consider two forecasts that are unbiased in absence of CLM, it can be seen from Proposition 2 that the memory of the loss differential is lower than that of the original series. The same holds true in presence of CLM, as covered by Proposition 5.

In practical situations, it might be overly restrictive to impose exact unbiasedness (under which memory would be reduced according to Proposition 5). Our empirical application regarding the predictive ability of the VIX serves as an example since it is a biased forecast of future quadratic variation due to the existence of a variance risk premium (see Section 6).

Biases can also be caused by estimation errors. This issue might be of less importance in a setup where the estimation period grows at a faster rate than the (pseudo-) out-of-sample period that is used for forecast evaluation. For the DM test, however, it is usually assumed that this is not the case. Otherwise, it could not be used for the comparison of forecasts from nested models due to a degenerated limiting distribution (cf. Giacomini and White (2006) for a discussion). Instead, the sample of size T^* is split into an estimation period T_E and a forecasting period T such that $T^* = T_E + T$ and it is assumed that T grows at a faster rate than T_E so that $T_E/T \to 0$ as $T^* \to \infty$. Therefore, the estimation error shrinks at a lower rate than the growth rate of the evaluation period and it remains relevant, asymptotically.

3.4 Asymptotic and Finite-Sample Behaviour under Long Memory

After establishing that forecast error loss differentials may exhibit long memory in various situations, we now consider the effect of long memory on the HAC-based Diebold-Mariano test. The following Proposition 5 establishes that the size of the test approaches unity, as $T \to \infty$. Thus, the test indicates with probability one that one of the forecasts is superior to the other one, even if both tests perform equally in terms of $g(\cdot)$.

Proposition 6 (DM under Long Memory)**.** *For $z_t \sim LM(d)$ with $d \in (0, 1/4) \cup (1/4, 1/2)$, the asymptotic size of the t_{HAC}-statistic equals unity as $T \to \infty$.*

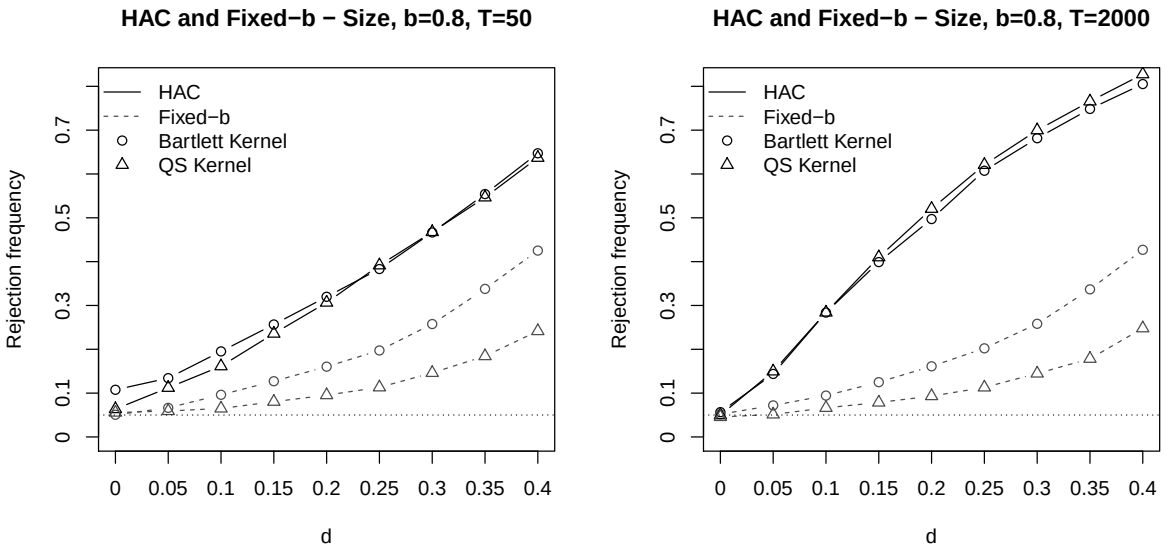

Figure 1: Size of the t_{HAC}- and t_{FB}-tests with $T \in \{50, 2000\}$ for different values of the memory parameter d.

Proof: *See the Appendix.*

This result shows that inference based on HAC estimators is asymptotically invalid under long memory. At the point $d = 1/4$, the asymptotic distribution of the t_{HAC}-statistic changes from normality to a Rosenblatt-type distribution which explains the discontinuity, see Abadir et al. (2009). In order to explore to what extent this finding also affects the finite-sample performance of the t_{HAC}- and t_{FB}-statistics, we conduct a small-scale Monte Carlo experiment as an illustration. The results shown in Figure 1 are obtained with $M = 5000$ Monte Carlo repetitions. We simulate samples of $T = 50$ and $T = 2000$ observations from a fractionally integrated process using different values of the memory parameter d in the range from 0 to 0.4. The HAC estimator and the fixed-b approach are implemented with the commonly used Bartlett- and Quadratic Spectral (QS) kernels.[2]

We start by commenting on the results for the small sample size of $T = 50$ in the left panel of Figure 1. As demonstrated by Kiefer and Vogelsang (2005), the fixed-b approach works exceptionally well for the short memory case of $d = 0$, with the Bartlett and QS kernel achieving approximately equal size control. The t_{HAC}-statistic behaves more liberal than the fixed-b approach and, as stated in Andrews (1991), better size control is provided if the Quadratic Spectral kernel is used. If the memory parameter d is positive, we observe that both tests severely over-reject the null hypothesis. For $d = 0.4$, the size of the HAC-based test is approximately 65% and that of the fixed-b version using the Bartlett kernel is around 40%. We therefore find that the size distortions are not only an asymptotic phenomenon, but they are already severe in samples of just $T = 50$ observations. Moreover, even for small deviations of d from zero, tests are over-sized. These findings motivate the use of long memory robust procedures. Continuing

[2]The bandwidth parameter of the fixed-b estimator is set to $b = 0.8$ since using a larger fraction of the autocorrelations provides a higher emphasis on size control (cf. Kiefer and Vogelsang (2005)). Other bandwidth choices lead to similar results.

with the results for $T = 2000$ in the right panel of Figure 1, we observe similar findings in general. For the short memory case, size distortions observed in small samples vanish. All tests statistics are well behaved for $d = 0$. On the contrary, for $d > 0$ size distortions are stronger compared to $T = 50$, although the magnitude of the additional distortion is moderate. This feature can be attributed to the slow divergence rate (as given in the proof of Proposition 6) of the test statistic under long memory.

4 Long-Run Variance Estimation under Long Memory

Since conventional HAC estimators lead to spurious rejections under long memory, we consider memory robust long-run variance estimators. To the best of our knowledge only two extensions of this kind are available in the literature: the memory and autocorrelation consistent (MAC) estimator of Robinson (2005) and an extension of the fixed-b estimator from McElroy and Politis (2012). We do not assume that forecasts are obtained from some specific class of model. We merely extend the typical assumptions of Diebold and Mariano (1995) on the loss differentials so that long memory is allowed.

4.1 MAC Estimator

The MAC estimator is developed by Robinson (2005) and further explored and extended by Abadir et al. (2009). Albeit stated in a somewhat different form, the same result is derived independently by Phillips and Kim (2007), who consider the long-run variance of a multivariate fractionally integrated process.

Robinson (2005) assumes that z_t is linear (in the sense of our equation (1), see also Assumption L in Abadir et al. (2009)) and that for $\lambda \to 0$ its spectral density fulfills

$$f(\lambda) = b_0|\lambda|^{-2d} + o(|\lambda|^{-2d}),$$

with $b_0 > 0$, $|\lambda| \leq \pi$, $d \in (-1/2, 1/2)$ and $b_0 = \lim_{\lambda \to 0} |\lambda|^{2d} f(\lambda)$.[3] Among others, this assumption covers stationary and invertible ARFIMA processes.

A key result for the MAC estimator is that as $T \to \infty$

$$\text{Var}\left(T^{1/2-d}\bar{z}\right) \to b_0 p(d)$$

with

$$p(d) = \begin{cases} \frac{2\Gamma(1-2d)\sin(\pi d)}{d(1+2d)} & \text{if } d \neq 0\,, \\ 2\pi & \text{if } d = 0\,. \end{cases}$$

The case of short memory ($d = 0$) yields the familiar result that the long-run variance of the sample mean equals $2\pi b_0 = 2\pi f(0)$. Hence, estimation of the long-run variance requires estimation of $f(0)$ in the case of short memory. If long memory is present in the data generating

[3] For notational convenience, here we drop the index z from the spectral density and the memory parameter.

process, estimation of the long-run variance additionally hinges on the estimation of d. The MAC estimator is therefore given by

$$\widehat{V}(\widehat{d}, m_d, m) = \widehat{b}_m(\widehat{d})p(\widehat{d}) \ .$$

In more detail, the estimation of V works as follows: First, if the estimator for d fulfills the condition $\widehat{d} - d = o_p(1/\log T)$, plug-in estimation is valid (cf. Abadir et al. (2009)). Thus, $p(d)$ can simply be estimated through $p(\widehat{d})$. A popular estimator that fulfills this rather weak requirement is the local Whittle estimator with bandwidth $m_d = \lfloor T^{q_d} \rfloor$, where $0 < q_d < 1$ denotes a generic bandwidth parameter and $\lfloor \cdot \rfloor$ denotes the largest integer smaller than its argument. This estimator is given by

$$\widehat{d}_{LW} = \arg \min_{d \in (-1/2, 1/2)} R_{LW}(d),$$

where $R_{LW}(d) = \log\left(\frac{1}{m_d}\sum_{j=1}^{m_d} j^{2d} I_T(\lambda_j)\right) - \frac{2d}{m_d}\sum_{j=1}^{m_d}\log j$, $I_T(\lambda_j)$ is the periodogram (which is independent of $\widehat{d}$),

$$I_T(\lambda_j) = (2\pi T)^{-1}\left|\sum_{t=1}^{T}\exp(\mathrm{i}t\lambda_j)z_t\right|^2$$

and the $\lambda_j = 2\pi j/T$ are the Fourier frequencies for $j = 1, ..., \lfloor T/2 \rfloor$.

Many other estimation approaches (e.g. log-periodogram estimation, etc.) would be a possibility as well. Since the loss differential in (7) is a linear combination of processes with different memory orders, the local polynomial Whittle plus noise (LPWN) estimator of Frederiksen et al. (2012) is a particularly useful alternative. This estimator extends the local Whittle estimator by approximating the log-spectrum of possible short memory components and perturbation terms in the vicinity of the origin by polynomials. This leads to a reduction of finite-sample bias. The estimator is consistent for $d \in (0, 1)$ and asymptotically normal in presence of perturbations for $d \in (0, 0.75)$, but with the variance inflated by a multiplicative constant compared to the local Whittle estimator.

Based on a consistent estimator $\widehat{d}$, as those discussed above, b_0 can be estimated consistently by

$$\widehat{b}_m(\widehat{d}) = m^{-1}\sum_{j=1}^{m}\lambda_j^{2\widehat{d}} I_T(\lambda_j).$$

The bandwidth m is determined according to $m = \lfloor T^q \rfloor$ such that $m \to \infty$ and $m = o(T/(\log T)^2)$. The MAC estimator is consistent as long as $\widehat{d} \xrightarrow{p} d$ and $\widehat{b}_m(\widehat{d}) \xrightarrow{p} b_0$. These results hold under very weak assumptions - neither linearity of z_t nor Gaussianity are required. Under somewhat stronger assumptions the t_{MAC}-statistic is also normal distributed (see Theorem 3.1. of Abadir et al. (2009)):

$$t_{MAC} \Rightarrow \mathcal{N}(0, 1) \ .$$

The t-statistic using the feasible MAC estimator can be written as

$$t_{MAC} = T^{1/2-\widehat{d}} \frac{\bar{z}}{\sqrt{\widehat{V}(\widehat{d}, m_d, m)}},$$

with m_d and m being the bandwidths for estimation of d and b_0, respectively.

It shall be noted that Abadir et al. (2009) also consider long memory versions of the classic HAC estimators. However, these extensions have two important shortcomings. First, asymptotic normality is lost for $1/4 < d < 1/2$ which complicates inference remarkably as d is generally unknown. Second, the extended HAC estimator is very sensitive towards the bandwidth choice as the MSE-optimal rate depends on d. On the contrary, the MAC estimator is shown to lead to asymptotically standard normally distributed t-ratios for the whole range of values $d \in (-1/2, 1/2)$. Moreover, the MSE-optimal bandwidth choice $m = \lfloor T^{4/5} \rfloor$ is independent of d. Thus, we focus on the MAC estimator and do not consider extended HAC estimators further.

4.2 Extended Fixed-Bandwidth Approach

Following up on the work by Kiefer and Vogelsang (2005), McElroy and Politis (2012) extend the fixed-bandwidth approach to long-range dependence. Their approach is similar to the one of Kiefer and Vogelsang (2005) in many respects, as can be seen below. The test statistic suggested by McElroy and Politis (2012) is given by

$$t_{EFB} = T^{1/2} \frac{\bar{z}}{\sqrt{\widehat{V}(k, b)}}.$$

In contrast to the t_{MAC}-statistic, the t_{EFB}-statistic involves a scaling of $T^{1/2}$. This has an effect on the limit distribution which depends on the memory parameter d. Analogously to the short memory case, the limiting distribution is derived by assuming that a functional central limit theorem for the partial sums of z_t applies, so that

$$t_{EFB} \Rightarrow \frac{W_d(1)}{\sqrt{Q(k, b, d)}},$$

where $W_d(r)$ is a fractional Brownian motion and $Q(k, b, d)$ depends on the fractional Brownian bridge $\widetilde{W}_d(r) = W_d(r) - rW_d(1)$. Furthermore, $Q(k, b, d)$ depends on the first and second derivatives of the kernel $k(\cdot)$. In more detail, for the *Bartlett* kernel we have

$$Q(k, b, d) = \frac{2}{b} \left(\int_0^1 \widetilde{W}_d(r)^2 \mathrm{d}r - \int_0^{1-b} \widetilde{W}_d(r+b) \widetilde{W}_d(r) \mathrm{d}r \right)$$

and thus, a similar structure as for the short memory case. Further details and examples can be found in McElroy and Politis (2012). The joint distribution of $W_d(1)$ and $\sqrt{Q(k, b, d)}$ is found through their joint Fourier-Laplace transformation, see Fitzsimmons and McElroy (2010). It is symmetric around zero and has a cumulative distribution function which is continuous in d.

Besides the similarities to the short memory case, there are some important conceptual differences to the MAC estimator. First, the MAC estimator belongs to the class of "small-b"

estimators in the sense that it estimates the long-run variance directly, whereas the fixed-b approach leads also in the long memory case to an estimate of the long-run variance multiplied by a functional of a *fractional* Brownian bridge. Second, the limiting distribution of the t_{EFB}-statistic is not a standard normal, but rather depending on the chosen kernel k, the fixed-bandwidth parameter b and the long memory parameter d. While the first two are user-specific, the latter one requires a plug-in estimator, as does the MAC estimator. As a consequence, the critical values are depending on d. McElroy and Politis (2012) offer response curves for various kernels.[4]

5 Monte Carlo Study

Further results on memory transmission to the forecast error loss differentials and the relative performance of the t_{MAC} and t_{EFB}-statistics are obtained by means of extensive Monte Carlo simulations. This section is divided into three parts. First, we conduct Monte Carlo experiments to verify the results obtained in Propositions 2 to 5 and to explore whether similar results apply for non-Gaussian processes and under the QLIKE loss function. The second part studies the memory properties of the loss differential in a number of empirically motivated forecasting scenarios. Finally, in the third part we explore the finite-sample size and power properties of the robustified tests discussed above and make recommendations for their practical application.

5.1 Memory Transmission to the Forecast Error Loss Differentials: Beyond MSE and Gaussianity

The results on the transmission of long memory from the forecasts or the forecast objective to the loss differentials in Propositions 2 to 5 are restricted to stationary Gaussian processes and forecasts evaluated using MSE as a loss function. In this section, we first verify the validity of the predictions from our propositions. Furthermore, we study how these results translate to non-Gaussian processes, non-stationary processes and the QLIKE loss function which we use in our empirical application in Section 6 on volatility forecasting. It is given by

$$QLIKE(\widehat{y}_{it}, y_t) = \log \widehat{y}_{it} + \frac{y_t}{\widehat{y}_{it}} \, . \tag{10}$$

For a discussion of the role and importance of this loss function in the evaluation of volatility forecasts see Patton (2011). All data generating processes are based on fractional integration. Due to the large number of cases in Propositions 2 to 5, we restrict ourselves to representative situations. The first two DGPs are based on cases (i) and (v) in Proposition 2 that covers situations when the forecasts and the forecast objective are generated from a system without common long memory. We simulate processes of the form

$$a_t = \mu_a + \frac{a_t^*}{\widehat{\sigma}_{a^*}}, \tag{11}$$

[4] All common kernels (e.g. Bartlett, Parzen) as well as others considered in Kiefer and Vogelsang (2005) can be used. In addition to the aforementioned, McElroy and Politis (2012) use the Daniell, the Trapezoid, the Modified Quadratic Spectral, the Tukey-Hanning and the Bohman kernel.

where $a_t \in \{y_t, \widehat{y}_{1t}, \widehat{y}_{2t}\}$, and $a_t^* = (1-L)^{-d_a}\varepsilon_{at}$. As in Section 3, the starred variable a_t^* is a zero-mean process, whereas a_t has mean μ_a and the ε_{at} are *iid*. The innovation sequences are either standard normal or $t(5)$-distributed. The standardization of a_t^* neutralizes the effect of increasing values of the memory parameter d on the process variance and controls the scaling of the mean relative to the variance. The loss differential series z_t is then calculated as in (1). We use 5000 Monte Carlo replications and consider sample sizes of $T = \{250, 2000\}$.

The first two DPGs for z_t are obtained by setting the means μ_a in (11) as follows

$$\begin{aligned}\text{DGP1: } &(\mu_1, \mu_2, \mu_y) = (1, -1, 0)\\ \text{DGP2: } &(\mu_1, \mu_2, \mu_y) = (0, 0, 0).\end{aligned}$$

The other DGPs represent the last cases of Propositions 3 to 5. These are based on the fractionally cointegrated system

$$\begin{pmatrix} y_t^* \\ \widehat{y}_{1t}^* \\ \widehat{y}_{2t}^* \\ x_t \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \xi_y \\ 0 & 1 & 0 & \xi_1 \\ 0 & 0 & 1 & \xi_2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_t \\ \varepsilon_{1t} \\ \varepsilon_{2t} \\ x_t \end{pmatrix},$$

where η_t, ε_{1t}, ε_{2t} and x_t are mutually independent and fractionally integrated with parameters d_η, d_{ε_1}, d_{ε_2} and d_x. DGPs 3 to 5 are then obtained by selecting the following parameter constellations:

$$\begin{aligned}\text{DGP3: } &(\mu_1, \mu_2, \mu_y, \xi_1, \xi_2, \xi_y) = (1, -1, 0, 1, 2, 1.5)\\ \text{DGP4: } &(\mu_1, \mu_2, \mu_y, \xi_1, \xi_2, \xi_y) = (1, -1, 0, 1, 1, 1)\\ \text{DGP5: } &(\mu_1, \mu_2, \mu_y, \xi_1, \xi_2, \xi_y) = (0, 0, 0, 1, 1, 1).\end{aligned}$$

Each of our DGPs 2 to 5 is formulated such that the reduction in the memory parameter is the strongest among all cases covered in the respective proposition. Simulation results for other cases would therefore show an even stronger transmission of memory to the loss differentials.

Since the QLIKE criterion is only defined for non-negative forecasts, we consider a long memory stochastic volatility specification if QLIKE is used and simulate forecasts and forecast objective of the form $\exp(a_t/2)$, whereas the MSE is calculated directly for the a_t.

It should be noted that the loss differential z_t is a linear combination of several persistent and antipersistent component series. This is a very challenging setup for the empirical estimation of the memory parameter. We therefore resort to the aforementioned LPWN estimator of Frederiksen et al. (2012) with a bandwidth of $m_d = \lfloor T^{0.65} \rfloor$ and a polynomial of degree one for the noise term that can be expected to have the lowest bias in this setup among the available methods to estimate the memory parameters. However, the estimation remains difficult and any mismatch between the theoretical predictions from our propositions and the finite-sample results reported here is likely to be due to the finite-sample bias of the semiparametric estimators.

The results for DGPs 1 and 2 are given in Table 1. We start with the discussion of simulation results for cases covered by our theoretical results. Table 1 shows the results for DGPs 1 and

			MSE								QLIKE							
			Gaussian				$t(5)$				Gaussian				$t(5)$			
DGP	T	d_1/d_2	0	0.2	0.4	0.6	0	0.2	0.4	0.6	0	0.2	0.4	0.6	0	0.2	0.4	0.6
1	250	0	0.32	0.32	0.35	0.43	0.30	0.31	0.34	0.41	0.29	0.32	0.38	0.44	0.22	0.25	0.32	0.41
		0.2	0.33	0.34	0.36	0.43	0.31	0.32	0.34	0.42	0.30	0.31	0.37	0.45	0.23	0.26	0.32	0.40
		0.4	0.36	0.36	0.38	0.45	0.33	0.34	0.37	0.43	0.31	0.32	0.37	0.45	0.24	0.27	0.33	0.41
		0.6	0.43	0.43	0.45	0.50	0.42	0.41	0.44	0.49	0.36	0.37	0.41	0.48	0.29	0.31	0.36	0.44
	2000	0	0.30	0.29	0.32	0.42	0.29	0.28	0.31	0.42	0.29	0.28	0.35	0.46	0.18	0.20	0.28	0.40
		0.2	0.29	0.29	0.32	0.41	0.28	0.28	0.31	0.41	0.29	0.28	0.35	0.46	0.18	0.20	0.27	0.40
		0.4	0.32	0.32	0.35	0.42	0.32	0.31	0.34	0.41	0.29	0.28	0.35	0.46	0.20	0.21	0.28	0.41
		0.6	0.42	0.42	0.43	0.48	0.42	0.41	0.42	0.47	0.35	0.34	0.38	0.48	0.26	0.25	0.31	0.44
2	250	0	0.13	0.15	0.26	0.43	0.10	0.14	0.23	0.41	0.11	0.14	0.26	0.41	0.10	0.13	0.20	0.35
		0.2	0.15	0.17	0.26	0.43	0.14	0.17	0.24	0.41	0.15	0.18	0.27	0.41	0.12	0.15	0.21	0.35
		0.4	0.26	0.27	0.31	0.43	0.23	0.24	0.29	0.42	0.25	0.27	0.31	0.41	0.20	0.21	0.25	0.36
		0.6	0.42	0.42	0.43	0.48	0.41	0.40	0.42	0.48	0.41	0.40	0.41	0.47	0.35	0.35	0.36	0.43
	2000	0	0.07	0.11	0.23	0.43	0.07	0.09	0.21	0.41	0.06	0.13	0.27	0.41	0.05	0.08	0.15	0.32
		0.2	0.11	0.13	0.23	0.42	0.09	0.11	0.20	0.40	0.13	0.17	0.26	0.40	0.07	0.09	0.17	0.31
		0.4	0.23	0.23	0.25	0.41	0.21	0.21	0.24	0.39	0.27	0.26	0.29	0.40	0.16	0.17	0.22	0.33
		0.6	0.43	0.42	0.41	0.46	0.41	0.40	0.39	0.45	0.41	0.41	0.40	0.46	0.32	0.31	0.33	0.41

Table 1: Monte Carlo averages of estimated memory in the loss differential z_t for DGP1 and DGP2 with $d_y = 0.25$.

2. Under MSE loss, and with Gaussian innovations, Proposition 2 states that for DGP1 we have $d_z = 0.25$ if $d_1, d_2 \in \{0, 0.2\}$ and $d_z = 0.4$ if either d_1 or d_2 is equal to 0.4, in the top left panel. In the bottom left panel results for DGP2 are reported. Proposition 2 states that $d_z = 0$ if $d_1, d_2 \in \{0, 0.2\}$ and $d_z = 0.3$, for $d_1 = 0.4$ or $d_2 = 0.4$. We can observe that the memory tends to be slightly larger than predicted for small d_1 and d_2 and it tends to be slightly smaller for $d_z = 0.4$. However, the results closely mirror the theoretical results from Proposition 2 in general.

With regard to the cases not covered by the theoretical derivations, we can observe that the results for t-distributed innovations are nearly identical to those obtained for the Gaussian distribution. The same holds true for the Gaussian long memory stochastic volatility model and the QLIKE loss function. If the innovations of the LMSV model are t-distributed, the memory in the loss differential is sligthly lower, but still substantial. Finally, in presence of non-stationary long memory with d_1 or d_2 equal to 0.6, we can observe that the loss differential exhibits long memory with an estimated degree between 0.4 and 0.5. The only exception is when the QLIKE loss function is used for DGP1. Here, we observe some asymmetry in the results, in the sense that the estimated memory parameter of the loss differential is slightly lower if d_2 is low, relative to d_1. However, the memory transmission is still substantial.

The results for DGP3 to DGP5, where forecasts and the forecast objective have common long memory is shown in Table 2. If we again consider the left column that displays the results for MSE loss and Gaussian innovations, Proposition 3 states for the case of DGP3 that the memory for all d in the stationary range should be $d_x = 0.45$. Proposition 4 does not give an exact prediction for DGP4, but states that the memory in the loss differential should be reduced compared to DGP3. Finally, for DGP5, Proposition 5 implies that $d_z = 0$, for $d_1, d_2 \in \{0, 0.2\}$ and $d_z = 0.3$ if d_1 or d_2 equal 0.4.

			MSE						QLIKE					
			Gaussian			$t(5)$			Gaussian			$t(5)$		
DGP	T	$d_{\varepsilon_1}/d_{\varepsilon_2}$	0	0.2	0.4	0	0.2	0.4	0	0.2	0.4	0	0.2	0.4
		0	0.31	0.32	0.38	0.29	0.31	0.37	0.29	0.33	0.40	0.27	0.31	0.39
	250	0.2	0.34	0.35	0.39	0.33	0.34	0.38	0.32	0.33	0.41	0.29	0.32	0.39
		0.4	0.41	0.42	0.44	0.40	0.40	0.43	0.37	0.38	0.43	0.36	0.37	0.42
3														
		0	0.34	0.32	0.38	0.34	0.32	0.38	0.32	0.30	0.39	0.31	0.29	0.38
	2000	0.2	0.30	0.30	0.36	0.30	0.30	0.35	0.30	0.29	0.38	0.29	0.29	0.37
		0.4	0.39	0.39	0.41	0.38	0.38	0.40	0.37	0.36	0.40	0.36	0.35	0.39
		0	0.29	0.31	0.35	0.27	0.30	0.35	0.28	0.30	0.37	0.24	0.27	0.35
	250	0.2	0.30	0.32	0.36	0.29	0.31	0.35	0.28	0.30	0.38	0.25	0.28	0.35
		0.4	0.35	0.36	0.39	0.34	0.35	0.38	0.31	0.32	0.39	0.27	0.30	0.37
4														
		0	0.26	0.26	0.33	0.25	0.26	0.33	0.25	0.25	0.35	0.23	0.24	0.33
	2000	0.2	0.26	0.26	0.33	0.26	0.25	0.32	0.26	0.26	0.35	0.24	0.24	0.33
		0.4	0.33	0.33	0.36	0.33	0.32	0.36	0.29	0.29	0.36	0.27	0.27	0.34
		0	0.12	0.14	0.25	0.11	0.13	0.23	0.10	0.13	0.25	0.10	0.12	0.21
	250	0.2	0.14	0.16	0.26	0.13	0.15	0.23	0.14	0.16	0.25	0.12	0.13	0.22
		0.4	0.26	0.25	0.30	0.23	0.24	0.28	0.25	0.25	0.30	0.22	0.22	0.26
5														
		0	0.06	0.10	0.23	0.07	0.10	0.21	0.05	0.10	0.25	0.04	0.07	0.19
	2000	0.2	0.09	0.11	0.23	0.09	0.11	0.21	0.09	0.12	0.25	0.06	0.09	0.19
		0.4	0.23	0.23	0.26	0.22	0.21	0.24	0.25	0.24	0.27	0.19	0.19	0.23

Table 2: Monte Carlo averages of estimated memory in the loss differential z_t for DGP3, DGP4 and DGP5 with $d_\eta = 0.2$.

As for DGP1 and DGP2, our estimates of d_z are roughly in line with the theoretical predictions. For DGP3, where d_z should be large, we see that the estimates are a bit lower, but the estimated degree of memory is still considerable. The results for DGP4 are indeed slightly lower than those for DGP3, as predicted by Proposition 4. Finally, for DGP5 we again observe that d_z is somewhat overestimated if the true value is low and vice versa. As in Table 1, we see that the results are qualitatively the same if we consider t-distributed innovation sequences and the QLIKE loss function. Additional simulations with $d_x = 0.65$ show that the results are virtually identical for $d_1, d_2 \leq 0.4$. If either $d_1 = 0.6$ or $d_2 = 0.6$, the memory transmission becomes even stronger, which is also in line with the findings for DGP1 and DGP2 in Table 1.

Overall, we find that the finite-sample results presented in this section are in line with the theoretical findings from Section 3. Moreover, the practical relevance of the results in Propositions 2 to 5 extends far beyond the stationary Gaussian case with MSE loss as demonstrated by the finding that the transmission results obtained with t-distributed innovations, non-stationary processes and the QLIKE loss function are nearly identical.

5.2 Forecast Scenarios

The relevance of memory transmission to the forecast error loss differentials in practice is further examined by considering a number of simple forecast scenarios motivated by typical empirical

T	$\widehat{d}_z$	$\bar{z}$	t_{HAC}	t_{MAC}	t_{EFB}
250	0.113	0.012	0.201	0.110	0.098
500	0.144	0.003	0.227	0.080	0.073
1000	0.171	0.000	0.270	0.057	0.058
2000	0.172	-0.001	0.324	0.046	0.057

Table 3: Estimated memory of the loss differentials $\widehat{d}_z$, mean loss differential $\bar{z}$, and rejection frequencies of the t-statistics for a spurious long memory scenario. The true DGP is fractionally integrated with random level shifts and the forecasts assume either a pure shift process or a pure long memory process.

examples. In order to ensure that the null hypothesis of equal predictive accuracy holds, we have to construct two competing forecasts that are different from each other, but perform equally in terms of a loss function - here the MSE. The length of the estimation period equals $T_E = 250$ and the memory parameter estimates are obtained by the LPWN estimator.

The first scenario is motivated by the spurious long memory literature. The DGP is a fractionally integrated process with a time-varying mean that is generated by a random level shift process as in Perron and Qu (2010) or Qu (2011). In detail,

$$\begin{aligned} y_t &= x_t + \mu_t \\ x_t &= (1-L)^{-1/4}\varepsilon_{x,t} \\ \mu_t &= \mu_{t-1} + \pi_t \varepsilon_{\mu,t}, \end{aligned}$$

where $\varepsilon_{x,t} \overset{iid}{\sim} N(0,1)$, $\varepsilon_{\mu,t} \overset{iid}{\sim} N(0,1)$, $\pi_t \overset{iid}{\sim} B(p)$ and ε_{xt}, $\varepsilon_{\mu t}$ and π_t are mutually independent.[5] It is well known that it can be difficult to distinguish long memory and low frequency contaminations such as structural breaks (cf. Diebold and Inoue (2001) or Granger and Hyung (2004)). Therefore, it is often assumed that the process is either driven by the one or the other, see e.g. Berkes et al. (2006), who suggest a test that allows to test for the null hypothesis of a weakly dependent process with breaks against the alternative of long-range dependence, or Lu and Perron (2010) who demonstrate that a pure level shift process has superior predictive performance compared to ARFIMA and HAR models for the log-absolute returns of the S&P 500. See also Varneskov and Perron (2017) for a related recent contribution.

In the spirit of this dichotomy, we compare forecasts which solely consider the breaks with those that assume the absence of breaks and predict the process based on a fractionally integrated model (with the memory estimated by the local Whittle method).[6]

Table 3 shows the results of this exercise. It is clear to see that the average loss differential is close to zero. The estimated memory of the loss differentials is around 0.17 for larger sample sizes. While the classical DM test based on a HAC estimator over-rejects, both the t_{MAC} and the t_{EFB}-statistics control the size well, at least in larger samples.

[5] To keep the implied degree of spurious long memory constant as the sample size increases, we set $p = 0.02$.

[6] To decrease the computational burden we take the actual mean shift process as a forecast. We thus abstract from estimation error in the (unconditional) mean component.

T	$\widehat{d}_z$	$\bar{z}$	t_{HAC}	t_{MAC}	t_{EFB}
250	0.182	0.011	0.289	0.063	0.059
500	0.207	-0.011	0.336	0.040	0.041
1000	0.228	-0.001	0.378	0.023	0.025
2000	0.238	-0.009	0.441	0.016	0.020

Table 4: Estimated memory of the loss differentials $\widehat{d}_z$, mean loss differential $\bar{z}$, and rejection frequencies of the t-statistics for comparison of forecasts obtained from predictive regressions where the regressor variables are fractionally cointegrated with the forecast objective.

As a second scenario, we consider simple predictive regressions based on two regressors that are fractionally cointegrated with the forecast objective. Here x_t is fractionally integrated of order d. Then

$$\begin{aligned} y_t &= x_t + (1-L)^{-(d-b)}\eta_t \\ x_{i,t} &= x_t + (1-L)^{-(d-b)}\varepsilon_{i,t} \\ \widehat{y}_{it} &= \widehat{\beta}_{0i} + \widehat{\beta}_{1i}x_{i,t-1}, \end{aligned}$$

where η_t and the $\varepsilon_{i,t}$ are mutually independent and normally distributed with unit variances, $\widehat{\beta}_{0i}$ and $\widehat{\beta}_{1i}$ are the OLS estimators and $0 < b < d$. To resemble processes in the lower non-stationary long memory region (as the realized volatilities in our empirical application) we set $d = 0.6$. This corresponds to a situation where we forecast realized volatility of the S&P 500 with either past values of the VIX or another past realized volatility such as that of a sector index. The cointegration strength is set to $b = 0.3$. The results are shown in Table 4. Again, one can see that the Monte Carlo averages ($\bar{z}$) are close to zero. The t_{MAC} and t_{EFB}-statistics tend to be conservative in larger samples, whereas the t_{HAC} test rejects far too often. The strength of the memory in the loss differential lies roughly at 0.24.

Our third scenario is closely related to the previous one. In practice it is hard to distinguish fractional cointegrated series from fractionally integrated series with highly correlated short-run components (cf. the simulation studies in Hualde and Velasco (2008)). Therefore, our third scenario is similar to the second, but with correlated innovations,

$$\begin{aligned} y_t &= (1-L)^{-d}\eta_t \\ x_{it} &= (1-L)^{-d}\varepsilon_{i,t} \\ \text{and} \qquad \widehat{y}_{it} &= \widehat{\beta}_{0i} + \widehat{\beta}_{1i}x_{i,t-1}. \end{aligned}$$

Here, all pairwise correlations between η_t and the $\varepsilon_{i,t}$ are $\rho = 0.4$. Furthermore, we set $d = 0.4$, so that we operate in the stationary long memory region. The situation is the same as in the previous scenarios, with strong long memory of approximately 0.3 in the loss differentials, see Table 5.

Altogether, these results demonstrate that memory transmission can indeed occur in a variety of situations whether that is due to level shifts, cointegration, or correlation - in a non-stationary

T	$\widehat{d}_z$	$\bar{z}$	t_{HAC}	t_{MAC}	t_{EFB}
250	0.275	0.001	0.364	0.028	0.026
500	0.288	0.001	0.417	0.017	0.020
1000	0.288	0.005	0.445	0.008	0.012
2000	0.287	0.003	0.485	0.004	0.009

Table 5: Estimated memory of the loss differentials $\widehat{d}_z$, mean loss differential $\bar{z}$, and rejection frequencies of the t-statistics for comparison of forecasts obtained from predictive regressions where the regressor variables have correlated innovations with the forecast objective.

series, or in a stationary series.

5.3 Finite-Sample Properties of Long Memory Robust t-Statistics

To explore the finite sample properties of the memory robust t_{EFB} and t_{MAC}-statistics, we conduct a number of size and power simulations. We use the same DGPs as in the previous sections in order to reflect the situations covered by our propositions and the distributional properties that are realistic for the forecast error loss differential z_t. However, we have to ensure that the loss differentials have zero expectation (size) and that the distance from the null is comparable for different DGPs (power).

As DGP1, DGP2, DGP4 and DGP5 are constructed in a symmetric way, we have $E[MSE(y_t, \widehat{y}_{1t}) - MSE(y_t, \widehat{y}_{2t})] = 0$ and $E[QLIKE(y_t, \widehat{y}_{1t}) - QLIKE(y_t, \widehat{y}_{2t})] \neq 0$, due to the asymmetry of the QLIKE loss function. Furthermore, DGP3 is not constructed in a symmetric way, so that $E[\tilde{z}_t] \neq 0$, irrespective of the loss function. We therefore have to correct the means of the loss differentials. In addition to that, different DGPs generate different degrees of long memory. Given that sample means of long memory processes with memory d converge with the rate $T^{1/2-d}$, we consider the local power to achieve comparable results across DGPs.

Let $\tilde{z}_t$ be generated as in (1), with the $\widehat{y}_{it}$ and y_t as described in Section 5.1, and $\bar{z} = (MT)^{-1}\sum_{i=1}^{M}\sum_{t=1}^{T}\tilde{z}_t$, where M denotes the number of Monte Carlo repetitions. Then the loss differentials are obtained via

$$z_t = \tilde{z}_t - \bar{z} + c\frac{SD(\tilde{z}_t)}{T^{1/2-d_z}}. \tag{12}$$

The parameter c controls the distance from the null hypothesis ($c = 0$). Here, each realization of $\tilde{z}_t$ is centered with the average sample mean from $M = 5000$ simulations of the respective DGP. Similarly, d_z is determined as the Monte Carlo average of the LPWN estimates for the respective setup. In the power simulations the memory parameters are set to $d_1 = d_2 = d$ and $d_{\varepsilon_1} = d_{\varepsilon_2} = d$ to keep the tables reasonably concise.

Table 6 presents the size results for the t_{MAC}-statistic. It tends to be liberal for small T, but generally controls the size well in larger samples. There are, however, two exceptions. First, the test remains liberal for DGP3, even if the sample size increases. This effect is particularly pronounced for $d = 0$ and if z_t is based on the QLIKE loss function. Second, the test is conservative for DGP2, particularly for increasing values of d. With regard to the bandwidth

		q	0.5						0.6					
			MSE			QLIKE			MSE			QLIKE		
q_d	T	DGP/d	0	0.2	0.4	0	0.2	0.4	0	0.2	0.4	0	0.2	0.4
		1	0.10	0.08	0.07	0.10	0.09	0.13	0.09	0.07	0.07	0.10	0.10	0.13
		2	0.01	0.02	0.03	0.02	0.04	0.05	0.01	0.03	0.03	0.02	0.04	0.04
	250	3	0.10	0.09	0.09	0.13	0.10	0.09	0.10	0.08	0.09	0.13	0.11	0.10
		4	0.10	0.08	0.09	0.10	0.09	0.10	0.09	0.09	0.09	0.10	0.08	0.10
		5	0.04	0.05	0.05	0.04	0.06	0.05	0.03	0.04	0.05	0.03	0.05	0.06
0.65														
		1	0.05	0.05	0.04	0.06	0.05	0.08	0.05	0.05	0.03	0.06	0.05	0.07
		2	0.02	0.02	0.01	0.02	0.05	0.02	0.01	0.02	0.01	0.02	0.04	0.02
	2000	3	0.07	0.05	0.04	0.15	0.11	0.06	0.06	0.05	0.04	0.15	0.11	0.05
		4	0.06	0.05	0.04	0.07	0.06	0.06	0.05	0.05	0.04	0.06	0.05	0.05
		5	0.04	0.05	0.02	0.04	0.06	0.02	0.03	0.05	0.02	0.03	0.06	0.02
		1	0.08	0.07	0.06	0.10	0.09	0.12	0.09	0.06	0.05	0.09	0.08	0.11
		2	0.02	0.02	0.02	0.02	0.03	0.04	0.02	0.02	0.02	0.02	0.04	0.04
	250	3	0.10	0.06	0.07	0.12	0.09	0.08	0.09	0.07	0.09	0.12	0.08	0.08
		4	0.09	0.07	0.08	0.10	0.08	0.10	0.08	0.06	0.07	0.10	0.08	0.10
		5	0.04	0.05	0.03	0.04	0.05	0.04	0.04	0.05	0.03	0.04	0.05	0.03
0.8														
		1	0.06	0.05	0.04	0.06	0.06	0.08	0.05	0.04	0.03	0.06	0.06	0.07
		2	0.02	0.01	0.00	0.02	0.04	0.02	0.02	0.01	0.00	0.02	0.04	0.02
	2000	3	0.07	0.05	0.06	0.16	0.11	0.07	0.06	0.04	0.05	0.14	0.11	0.06
		4	0.06	0.04	0.06	0.06	0.05	0.07	0.05	0.04	0.04	0.06	0.05	0.06
		5	0.04	0.04	0.01	0.04	0.05	0.02	0.04	0.04	0.01	0.04	0.05	0.01

Table 6: Size results of the t_{MAC}-statistic for the DGP described in Section 5.1 and (12) with Gaussian innovations.

parameters, we find that the size is slightly better controlled with $q_d = 0.65$ and $q = 0.6$. However, the bandwidth choice seems to have limited effects on the size of the test, especially in larger samples.

Size results for the t_{EFB}-statistic are displayed in Table 7. To analyze the impact of the bandwidth b and the kernel choice, we set $q_d = 0.65$.[7] Size performance is more favorable with the MQS kernel than with the Bartlett kernel. Furthermore, it is positively impacted by using a large value of b (0.6 or 0.9). Similar to the t_{MAC}-statistic, we observe that the test is liberal with $T = 250$, but that the overall performance is very satisfactory for $T = 2000$. Again, the test tends to be liberal for DGP3 - especially for QLIKE. However, if the MQS kernel and a larger b is used, this effect disappears nearly completely. The conservative behavior of the test for DGP2 and large values of d is also the same as for the t_{MAC}-statistic. The t_{MAC}-statistic tends to be perform better than the t_{EFB}-statistic using the Bartlett kernel, but worse when considering the MQS kernel (cf. Tables 6 and 7).

With regard to the power results in Table 8, we find that the power tends to be a bit lower for $d = 0.4$, which is likely due to the fact that the scaling parameter d_z in (12) tends to be underestimated if the true value is larger (cf. Section 5.1). In general, the power of the t_{EFB}-statistic appears to be better if the Bartlett kernel is used compared to the MQS kernel. For

[7] Additional simulations with $q_d = 0.80$ yield nearly identical results.

			MSE						QLIKE					
		kernel	MQS			Bartlett			MQS			Bartlett		
d	T	DGP/b	0.3	0.6	0.9	0.3	0.6	0.9	0.3	0.6	0.9	0.3	0.6	0.9
		1	0.08	0.06	0.06	0.10	0.10	0.08	0.07	0.06	0.06	0.10	0.10	0.09
		2	0.03	0.03	0.03	0.02	0.02	0.02	0.04	0.03	0.03	0.02	0.02	0.02
	250	3	0.12	0.10	0.09	0.16	0.14	0.14	0.11	0.09	0.08	0.15	0.14	0.14
		4	0.07	0.06	0.06	0.10	0.09	0.09	0.08	0.06	0.07	0.10	0.10	0.09
		5	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.05	0.04	0.04	0.04
0														
		1	0.06	0.05	0.05	0.07	0.05	0.06	0.06	0.04	0.05	0.07	0.06	0.06
		2	0.03	0.03	0.03	0.02	0.02	0.03	0.04	0.03	0.03	0.03	0.03	0.03
	2000	3	0.07	0.06	0.06	0.10	0.10	0.09	0.08	0.06	0.07	0.10	0.09	0.10
		4	0.06	0.05	0.05	0.06	0.06	0.06	0.07	0.06	0.06	0.08	0.07	0.07
		5	0.04	0.04	0.04	0.05	0.05	0.04	0.05	0.04	0.04	0.05	0.05	0.05
		1	0.06	0.05	0.05	0.08	0.08	0.07	0.08	0.06	0.06	0.09	0.08	0.08
		2	0.04	0.02	0.03	0.03	0.03	0.03	0.04	0.03	0.03	0.04	0.04	0.04
	250	3	0.09	0.07	0.08	0.11	0.10	0.10	0.09	0.07	0.07	0.11	0.10	0.10
		4	0.07	0.06	0.06	0.08	0.08	0.07	0.07	0.06	0.05	0.09	0.09	0.08
		5	0.05	0.03	0.04	0.04	0.05	0.05	0.05	0.05	0.04	0.05	0.05	0.06
0.2														
		1	0.05	0.04	0.04	0.06	0.05	0.05	0.05	0.04	0.04	0.07	0.05	0.05
		2	0.03	0.03	0.03	0.02	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.05
	2000	3	0.07	0.05	0.06	0.08	0.07	0.08	0.07	0.06	0.06	0.07	0.07	0.08
		4	0.05	0.04	0.04	0.06	0.06	0.05	0.06	0.05	0.05	0.06	0.06	0.06
		5	0.06	0.04	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.07	0.07	0.06
		1	0.05	0.05	0.05	0.07	0.07	0.06	0.10	0.09	0.08	0.14	0.13	0.11
		2	0.03	0.03	0.03	0.03	0.03	0.04	0.05	0.04	0.04	0.05	0.04	0.05
	250	3	0.08	0.07	0.07	0.09	0.09	0.08	0.06	0.06	0.05	0.09	0.08	0.08
		4	0.08	0.06	0.07	0.09	0.09	0.09	0.09	0.07	0.06	0.11	0.11	0.11
		5	0.04	0.03	0.03	0.05	0.05	0.05	0.04	0.04	0.04	0.05	0.05	0.05
0.4														
		1	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.05	0.06	0.08	0.08	0.07
		2	0.02	0.02	0.02	0.02	0.02	0.01	0.03	0.03	0.03	0.03	0.03	0.03
	2000	3	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
		4	0.05	0.05	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06
		5	0.03	0.03	0.03	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03

Table 7: Size results of the t_{EFB}-statistic for the DGP described in Section 5.1 and (12) with Gaussian innovations and $m = \lfloor T^{0.65} \rfloor$.

DGP1, the power of the test using the MQS kernel is particularly low. The performance of the t_{MAC}-statistic is somewhere in between the two t_{EFB}-statistics which means that the ordering is directly inverse to the performance in terms of size control. It can be observed in Table 8 that, given a specific value of the loss differential, tests using the QLIKE loss function are generally more powerful compared to tests using MSE loss.[8]

Altogether, we find that the t_{EFB}-statistic using the MQS kernel should be used in combination with the QLIKE loss function and for smaller samples since it is the only specification that offers reliable size control. In larger samples and under MSE loss all of the tests have satisfactory

[8]This is in line with short memory simulations in Patton and Sheppard (2009).

			t_{EFB}																t_{MAC}							
			MSE								QLIKE								MSE				QLIKE			
		kernel	Bartlett				MQS				Bartlett				MQS											
d	T	DGP/c	0	3	6	9	0	3	6	9	0	3	6	9	0	3	6	9	0	3	6	9	0	3	6	9
		1	0.09	0.15	0.31	0.45	0.06	0.10	0.19	0.30	0.10	0.69	0.93	0.99	0.06	0.56	0.79	0.91	0.10	0.17	0.34	0.49	0.11	0.65	0.72	0.72
		2	0.02	0.25	0.68	0.83	0.03	0.13	0.37	0.59	0.02	0.94	1.00	1.00	0.03	0.89	0.96	0.99	0.01	0.29	0.76	0.84	0.02	0.88	0.90	0.90
	250	3	0.10	0.46	0.70	0.86	0.07	0.31	0.58	0.72	0.13	1.00	1.00	1.00	0.08	0.96	1.00	1.00	0.11	0.50	0.67	0.72	0.14	0.76	0.76	0.78
		4	0.09	0.28	0.54	0.68	0.06	0.17	0.38	0.55	0.10	0.95	1.00	1.00	0.06	0.83	0.98	1.00	0.10	0.31	0.57	0.67	0.12	0.75	0.75	0.76
		5	0.05	0.69	0.87	0.92	0.03	0.40	0.76	0.86	0.05	1.00	1.00	1.00	0.04	0.98	1.00	1.00	0.04	0.75	0.87	0.87	0.04	0.90	0.91	0.91
0																										
		1	0.06	0.13	0.30	0.49	0.04	0.09	0.18	0.32	0.06	0.78	0.97	1.00	0.05	0.60	0.86	0.95	0.06	0.15	0.34	0.52	0.07	0.69	0.79	0.81
		2	0.02	0.44	0.84	0.90	0.03	0.21	0.53	0.76	0.03	0.99	1.00	1.00	0.03	0.96	1.00	1.00	0.01	0.53	0.88	0.89	0.02	0.94	0.95	0.95
	2000	3	0.07	0.57	0.87	0.96	0.05	0.38	0.71	0.85	0.15	1.00	1.00	1.00	0.09	0.99	1.00	1.00	0.07	0.61	0.79	0.84	0.16	0.90	0.91	0.92
		4	0.06	0.32	0.66	0.81	0.05	0.20	0.46	0.66	0.08	0.99	1.00	1.00	0.05	0.92	1.00	1.00	0.06	0.36	0.69	0.76	0.07	0.85	0.88	0.89
		5	0.04	0.88	0.94	0.98	0.04	0.62	0.90	0.95	0.05	1.00	1.00	1.00	0.04	1.00	1.00	1.00	0.03	0.90	0.92	0.92	0.03	0.96	0.96	0.96
		1	0.08	0.13	0.24	0.38	0.05	0.09	0.16	0.26	0.10	0.62	0.88	0.97	0.06	0.49	0.74	0.87	0.08	0.15	0.28	0.42	0.11	0.59	0.69	0.7
		2	0.03	0.18	0.49	0.68	0.03	0.10	0.27	0.46	0.04	0.86	0.98	1.00	0.03	0.76	0.92	0.98	0.02	0.20	0.55	0.73	0.03	0.81	0.83	0.84
	250	3	0.09	0.35	0.62	0.79	0.06	0.22	0.48	0.63	0.11	1.00	1.00	1.00	0.07	0.93	1.00	1.00	0.09	0.38	0.63	0.70	0.12	0.76	0.78	0.76
		4	0.08	0.21	0.44	0.61	0.05	0.14	0.31	0.46	0.10	0.89	1.00	1.00	0.05	0.76	0.95	0.99	0.09	0.23	0.48	0.62	0.10	0.72	0.75	0.76
		5	0.05	0.56	0.79	0.86	0.04	0.34	0.64	0.79	0.05	0.99	1.00	1.00	0.04	0.94	1.00	1.00	0.04	0.62	0.79	0.83	0.05	0.85	0.87	0.87
0.2																										
		1	0.05	0.11	0.26	0.44	0.04	0.08	0.16	0.27	0.06	0.74	0.95	0.99	0.05	0.54	0.83	0.93	0.05	0.13	0.29	0.51	0.06	0.72	0.84	0.87
		2	0.03	0.26	0.64	0.80	0.03	0.14	0.37	0.59	0.05	0.91	1.00	1.00	0.04	0.82	0.95	0.99	0.02	0.29	0.71	0.81	0.04	0.83	0.86	0.87
	2000	3	0.06	0.47	0.82	0.93	0.05	0.30	0.62	0.80	0.11	1.00	1.00	1.00	0.08	0.98	1.00	1.00	0.05	0.54	0.81	0.87	0.11	0.92	0.94	0.94
		4	0.06	0.25	0.58	0.76	0.05	0.16	0.36	0.56	0.06	0.97	1.00	1.00	0.04	0.88	0.99	1.00	0.05	0.27	0.64	0.78	0.05	0.88	0.92	0.93
		5	0.06	0.73	0.89	0.94	0.05	0.47	0.79	0.89	0.07	1.00	1.00	1.00	0.05	0.99	1.00	1.00	0.05	0.78	0.86	0.88	0.07	0.90	0.91	0.91
		1	0.07	0.10	0.17	0.25	0.05	0.08	0.12	0.18	0.12	0.47	0.72	0.86	0.10	0.35	0.57	0.71	0.07	0.11	0.19	0.28	0.14	0.43	0.59	0.65
		2	0.03	0.08	0.18	0.29	0.03	0.05	0.11	0.19	0.05	0.59	0.83	0.93	0.03	0.46	0.68	0.81	0.03	0.08	0.21	0.31	0.05	0.59	0.69	0.72
	250	3	0.10	0.17	0.37	0.55	0.07	0.13	0.26	0.39	0.10	0.93	1.00	1.00	0.07	0.78	0.96	1.00	0.11	0.19	0.40	0.52	0.11	0.67	0.69	0.70
		4	0.08	0.14	0.25	0.37	0.07	0.11	0.18	0.25	0.10	0.67	0.91	0.98	0.08	0.52	0.77	0.90	0.10	0.16	0.27	0.38	0.11	0.57	0.66	0.68
		5	0.04	0.19	0.39	0.51	0.04	0.12	0.27	0.39	0.05	0.82	0.96	0.99	0.04	0.70	0.90	0.95	0.05	0.21	0.41	0.54	0.06	0.70	0.76	0.77
0.4																										
		1	0.05	0.07	0.12	0.21	0.03	0.05	0.09	0.14	0.07	0.43	0.71	0.87	0.06	0.30	0.54	0.71	0.04	0.07	0.13	0.24	0.08	0.45	0.68	0.78
		2	0.01	0.04	0.12	0.24	0.02	0.04	0.09	0.16	0.03	0.56	0.81	0.92	0.03	0.41	0.67	0.79	0.01	0.02	0.11	0.24	0.03	0.60	0.78	0.84
	2000	3	0.05	0.12	0.32	0.54	0.05	0.08	0.21	0.35	0.06	0.94	1.00	1.00	0.05	0.78	0.97	1.00	0.05	0.12	0.38	0.60	0.06	0.82	0.87	0.87
		4	0.06	0.09	0.17	0.31	0.04	0.07	0.12	0.20	0.06	0.63	0.91	0.98	0.05	0.46	0.76	0.89	0.05	0.10	0.20	0.37	0.06	0.64	0.81	0.84
		5	0.03	0.13	0.35	0.54	0.03	0.09	0.21	0.36	0.03	0.83	0.97	0.99	0.03	0.69	0.88	0.95	0.02	0.13	0.38	0.57	0.02	0.80	0.88	0.90

Table 8: Local power for both the t_{EFB} and t_{MAC}-statistic for the DGP described in Section 5.1 and (12) with Gaussian innovations, $b = 0.6$, and $m = \lfloor T^{0.65} \rfloor$.

size. In this case, the t_{EFB}-statistic with the Bartlett kernel is preferable in terms of power. However, since the size control of the t_{MAC}-statistic is better than that of the t_{EFB}-statistic with the Bartlett kernel, and the power is better than that using the MQS kernel, it can still be a sensible option in intermediate cases.

6 Applications to Realized Volatility Forecasting

Due to its relevance for risk management and derivative pricing, volatility forecasting is of vital importance and is also one of the fields in which long memory models are applied most often (cf. Deo et al. (2006), Martens et al. (2009) and Chiriac and Voev (2011)). Since intraday data on financial transactions has become widely available, the focus has shifted from GARCH-type models to the direct modelling of realized volatility series. In particular the heterogeneous autoregressive model (HAR-RV) of Corsi (2009) and its extensions have emerged as one of the most popular approaches.

We re-evaluate some recent results from the related literature using traditional Diebold-Mariano tests as well as the long memory robust versions from Section 4. We use a data set of 5-minute log-returns of the S&P 500 Index from January 2, 1996 to August 31, 2015 and we include close-to-open returns. In total, we have $T = 4883$ observations in our sample. The raw data is obtained from the Thomson Reuters Tick History Database.

Before we turn to the forecast evaluations in Sections 6.1 and 6.2, we use the remainder of this section to define the relevant volatility variables and to introduce the data and the employed time series models. Define the j-th intraday return on day t by $r_{t,j}$ and let there be N intraday returns per day. The daily realized variance is then defined as

$$RV_t = \sum_{j=1}^{N} r_{t,j}^2 \ ,$$

see e.g. Andersen et al. (2001) and Barndorff-Nielsen and Shephard (2002). If $r_{t,j}$ is sampled with an ever-increasing frequency such that $N \to \infty$, RV_t provides a consistent estimate of the quadratic variation of the log-price process. Therefore, RV_t is usually treated as a direct observation of the stochastic volatility process. The HAR-RV model of Corsi (2009), for example, explains log-realized variance by an autoregression involving overlapping averages of past realized variances. Similar to the notation in Bekaert and Hoerova (2014), the model reads

$$\ln RV_t^{(h)} = \alpha + \rho_{22} \ln RV_{t-h}^{(22)} + \rho_5 \ln RV_{t-h}^{(5)} + \rho_1 \ln RV_{t-h}^{(1)} + \varepsilon_t \ , \tag{13}$$

where $RV_t^{(M)} = \frac{22}{M} \sum_{j=0}^{M-1} RV_{t-j}$ and ε_t is a white noise process. Although this is formally not a long memory model, this simple process provides a good approximation to the slowly decaying autocorrelation functions of long memory processes in finite samples. Forecast comparisons show that the HAR-RV model performs similar to ARFIMA models (cf. Corsi (2009)).

Motivated by developments in derivative pricing highlighting the importance of jumps in price processes, Andersen et al. (2007) extend the HAR-RV model to consider jump components in realized volatility. Here, the underlying model for the continuous time log-price process $p(t)$ is

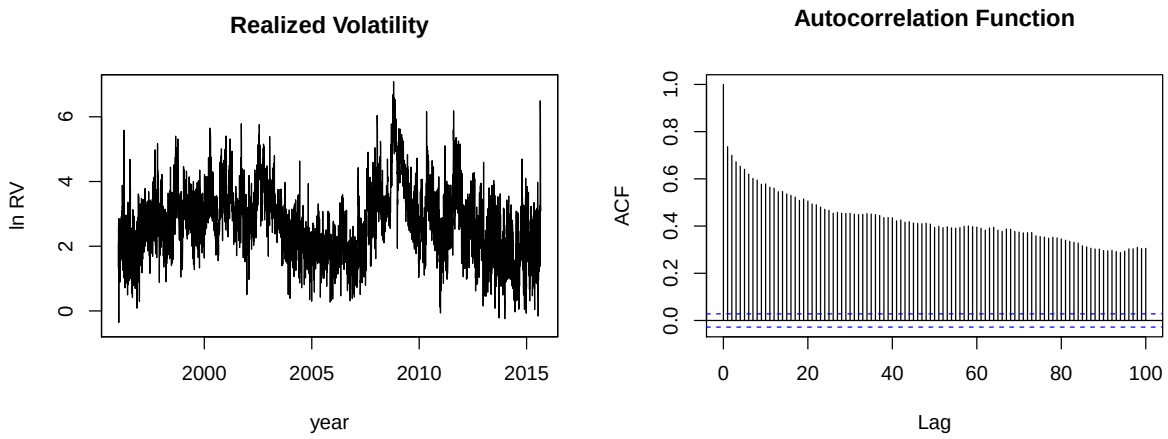

Figure 2: Daily log-realized volatility of the S&P500 index and their autocorrelation function.

given by

$$dp(t) = \mu(t)dt + \sigma(t)dW(t) + \kappa(t)dq(t) \ ,$$

where $0 \leq t \leq T$, $\mu(t)$ has locally bounded variation, $\sigma(t)$ is a strictly positive stochastic volatility process that is càdlàg and $W(t)$ is a standard Brownian motion. The counting process $q(t)$ takes the value $dq(t) = 1$ if a jump is realized and it is allowed to have time-varying intensity. Finally, the process $\kappa(t)$ determines the size of discrete jumps in case these are realized. The quadratic variation of the cumulative return process can be thus decomposed into integrated volatility plus the sum of squared jumps:

$$[r]_t^{t+h} = \int_t^{t+h} \sigma^2(s)ds + \sum_{t<s\leq t+h} \kappa^2(s) \ .$$

In order to measure the integrated volatility component, Barndorff-Nielsen and Shephard (2004, 2006) introduce the concept of bipower variation (BPV) as an alternative estimator that is robust to the presence of jumps. Here, we use threshold bipower variation (TBPV) as suggested by Corsi et al. (2010), who showed that BPV can be severely biased in finite samples. TBPV is defined as follows:

$$TBPV_t = \frac{\pi}{2} \sum_{j=2}^{N} |r_{t,j}||r_{t,j-1}| \mathbb{I}(|r_{t,j}|^2 \leq \zeta_j) \mathbb{I}(|r_{t,j-1}|^2 \leq \zeta_{j-1}) \ ,$$

where ζ_j is a strictly positive, random threshold function as specified in Corsi et al. (2010) and $\mathbb{I}(\cdot)$ is an indicator function.[9] Since

$$TBPV_t \xrightarrow{p} \int_t^{t+1} \sigma^2(s)ds$$

for $N \to \infty$, one can decompose the realized volatility into the continuous integrated volatility

[9] To calculate ζ_j, we closely follow Corsi et al. (2010).

q	$\widehat{d}_{LW}$	$\widehat{d}_{HP}$	s.e.	$\widetilde{W}_z$	$\widehat{d}_{(0,0)}$		$\widehat{d}_{(1,0)}$		$\widehat{d}_{(1,1)}$	
0.55	0.554	0.493	(0.048)	0.438	0.613	(0.088)	0.612	(0.132)	0.689	(0.163)
0.60	0.553	0.522	(0.039)	0.568	0.567	(0.074)	0.577	(0.110)	0.692	(0.131)
0.65	0.573	0.573	(0.032)	0.544	0.573	(0.059)	0.570	(0.089)	0.570	(0.118)
0.70	0.549	0.532	(0.026)	0.449	0.573	(0.048)	0.578	(0.072)	0.588	(0.093)
0.75	0.539	0.518	(0.021)	0.515	0.564	(0.039)	0.574	(0.058)	0.593	(0.075)

Table 9: Long memory estimation and testing results for S&P 500 log-realized volatility. Local Whittle estimates for the d parameter and results of the Qu (2011) test ($\widetilde{W}_z$ modified statistic by Kruse (2015)) for true versus spurious long memory are reported for various bandwidth choices $m_d = \lfloor T^q \rfloor$. Critical values are 1.118, 1.252 and 1.517 at the nominal significance level of 10%, 5% and 1%, respectively. Asymptotic standard errors for $\widehat{d}_{LW}$ and $\widehat{d}_{HP}$ are given in parentheses. The indices of the LPWN estimators indicate the orders of the polynomials used.

component C_t and the jump component J_t as

$$J_t = \max\{RV_t - TBPV_t, 0\}\, \mathbb{I}(\text{C-Tz} > 3.09)\ ,$$
$$C_t = RV_t - J_t\ .$$

The argument of the indicator function $\mathbb{I}(\text{C-Tz} > 3.09)$ ensures that the jump component is set to zero if it is insignificant at the nominal 0.1% level, so that J_t is not contaminated by measurement error, see also Corsi and Renò (2012). For details on the C-Tz-statistic, see Corsi et al. (2010).

Different from previous studies reporting an insignificant or negative impact of jumps, Corsi et al. (2010) show that the impact of jumps on future realized volatility is significant and positive. We use the HAR-RV-TCJ model that is studied in Bekaert and Hoerova (2014):

$$\begin{aligned}\ln RV_t^{(h)} = {} & \alpha + \rho_{22} \ln C_{t-h}^{(22)} + \rho_5 \ln C_{t-h}^{(5)} + \rho_1 \ln C_{t-h}^{(1)} \\ & + \varpi_{22}\ \ln\left(1 + J_{t-h}^{(22)}\right) + \varpi_5 \ln\left(1 + J_{t-h}^{(5)}\right) + \varpi_1 \ln\left(1 + J_{t-h}^{(1)}\right) + \varepsilon_t\ . \end{aligned} \tag{14}$$

The daily log-realized variance series ($\ln RV_t$) is depicted in Figure 2.[10] It is common to use log-realized variance to avoid non-negativity constraints on the parameters and to have a better approximation to the normal distribution, as advocated by Andersen et al. (2001). As can be seen from Figure 2, the series shows the typical features of a long memory time series, namely a hyperbolically decaying autocorrelation function, as well as local trends.

Estimates of the memory parameter are shown in Table 9. Local Whittle estimates ($\widehat{d}_{LW}$) exceed 0.5 slightly and thus indicate a mild form of non-stationarity. Since there is a large literature on the potential of spurious long memory in volatility time series, we carry out the test of Qu (2011). To avoid issues due to non-stationarity and to increase the power of the test, we follow Kruse (2015) and apply the test to the fractional difference of the data. The necessary degree of differencing is determined using the estimator by Hou and Perron (2014) ($\widehat{d}_{HP}$) that is robust

[10] For a better comparison, all variables in this section are scaled towards a monthly basis.

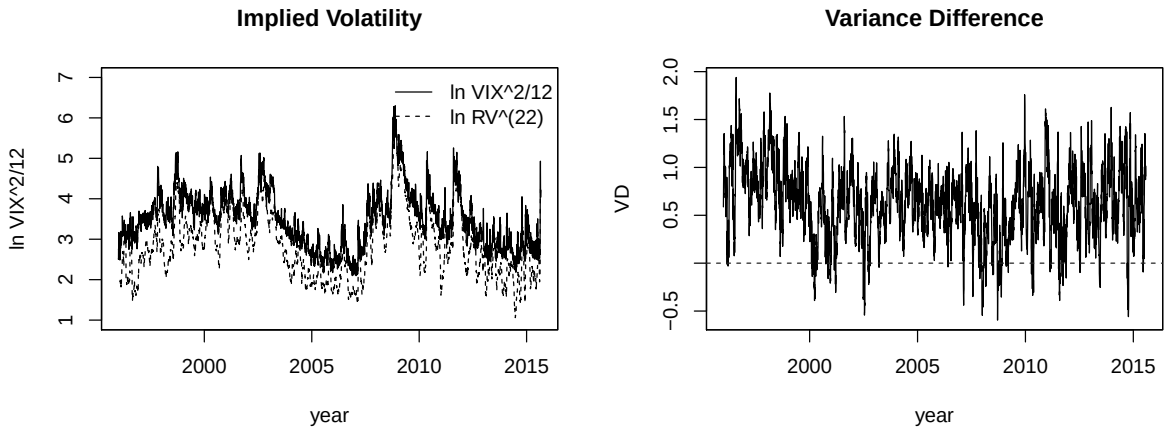

Figure 3: Log squared implied volatility and log cumulative realized volatility of the S&P 500 (left panel) and variance difference $VD_t = \ln(VIX_t^2/12) - \ln RV_{t+22}^{(22)}$ (right panel).

to low-frequency contaminations. A comparison of the $\widetilde{W}_z$ statistic to its critical values reveal that the test fails to reject the null hypothesis of true long memory.

Since N is finite in practice, RV_t might contain a measurement error and is therefore often modeled as the sum of the quadratic variation and an *iid* perturbation process such that $RV_t = [r]_t^{t+1} + u_t$, where $u_t \sim iid(0, \sigma_u^2)$. Furthermore, it is well known that local Whittle estimates can be biased in presence of short-run dynamics. To this end, we report results of LPWN estimator. The estimates remain remarkably stable - irrespective of the choice of the estimator. The downward bias of the local Whittle estimator due to the measurement error in realized variance is therefore moderate.

Altogether, the realized variance series appears to be a long memory process. Consequently, if forecasts of the series are evaluated, a transmission of long-range dependence to the loss differentials as implied by Propositions 2 to 5 may occur. This would invalidate conventional DM tests, as shown in Proposition 6 and Figure 1 highlighting the importance of the robust t_{MAC} and t_{EFB}-statistics discussed in Section 4.

6.1 Predictive Ability of the VIX for Quadratic Variation

The predictive ability of implied volatility for future realized volatility is an issue that has received a lot of attention in the related literature. The CBOE VIX represents the market expectation of quadratic variation of the S&P 500 over the next month, derived under the assumption of risk neutral pricing. Both, $\ln(VIX_t^2/12)$ and $\ln RV_{t+22}^{(22)}$ are depicted in Figure 3. Both series behave fairly similar and are quite persistent. As for the log-realized volatility series, the Qu (2011) test does not reject the null hypothesis of true long memory for the VIX after appropriate fractional differencing following Kruse (2015).

Chernov (2007) investigates the role of a variance risk premium in the market for volatility forecasting. The variance risk premium is given by $VP_t = VIX_t^2/12 - RV_{t+22}^{(22)}$. A related variable that is used for example by Bollerslev et al. (2013) is the variance difference $VD_t = \log(VIX_t^2/12) - \log(RV_{t+22}^{(22)})$ that is displayed on the right hand side of Figure 3. The graph

	Summary statistics					Short memory inference			Long memory inference						
									t_{MAC}			t_{EFB}			
Model	$\overline{z}/\widehat{\sigma}_z$	g_1	g_2	$\widehat{d}_{LW}$	$\widehat{d}_{LPWN}$	t_{DM}	t_{HAC}	t_{FB}	0.7	0.75	0.8	0.2	0.4	0.6	0.8
HAR-RV	0.14	0.29	0.27	**0.22***	**0.24***	**2.97**	**3.03**	**2.49**	0.93	1.04	1.18	2.49 (3.40)	2.75 (4.06)	2.99 (4.75)	2.85 (5.39)
HAR-RV-TCJ	0.11	0.26	0.27	**0.18***	0.14	**2.42**	**2.46**	**2.10**	1.40	1.61	**1.89**	2.10 (2.61)	2.50 (3.15)	2.89 (3.69)	2.72 (4.23)
HAR-RV-TCJ-L	0.08	0.28	0.27	**0.18***	0.16	**1.78** (1.65)	**1.79** (1.65)	1.82 (2.09)	0.90	1.03 (1.65)	1.20	1.82 (3.40)	2.15 (4.06)	2.43 (4.75)	2.32 (5.39)

Table 10: Predictive ability of the VIX for future RV (evaluated under MSE loss). Models excluding the VIX are tested against models including the VIX. Reported are the standardized mean ($\overline{z}/\widehat{\sigma}_z$) and estimated memory parameter ($\widehat{d}$) of the forecast error loss differential. Furthermore, the respective out-of-sample loss of the models (g_1 and g_2) and the results of various DM test statistics are given. Bold-faced values indicate significance at the nominal 5% level; an additional star indicates significance at the nominal 1% level. Critical values of the tests are given in parentheses.

clearly suggests that the VIX tends to overestimate the realized variance and the sample average of the variance difference is 0.623. Furthermore, the linear combination of log-realized and log-implied volatility is rather persistent and has a significant memory of $\widehat{d}_{LPWN} = 0.2$. This is consistent with the existence of a fractional cointegration relationship between $\ln(VIX_t^2/12)$ and $\ln RV_{t+22}^{(22)}$ which has been considered in several contributions including Christensen and Nielsen (2006), Nielsen (2007) and Bollerslev et al. (2013). Bollerslev et al. (2009), Bekaert and Hoerova (2014) and Bollerslev et al. (2013) additionally extend the analysis towards the predictive ability of VD_t for stock returns.

While the aforementioned articles test the predictive ability of the VIX itself and the "implied-realized-parity", there has also been a series of studies that analyze whether the inclusion of implied volatility can improve model-based forecasts. On the one hand, Becker et al. (2007) conclude that the VIX does not contain any incremental information on future volatility relative to an array of forecasting models. On the other hand, Becker et al. (2009) show that the VIX is found to subsume information on past jump activity and contains incremental information on future jumps if continuous components and jump components are considered separately. Similarly, Busch et al. (2011) study a HAR-RV model with continuous components and jumps and propose a VecHAR-RV model. They find that the VIX has incremental information and partially predicts jumps.

Motivated by these findings, we test whether the inclusion of $\ln(VIX_t^2/12)$ improves model-based forecasts from HAR-RV-type models, using Diebold-Mariano statistics. Since the VIX can be seen as a forecast of future quadratic variation over the next month, we consider a 22-step forecast horizon. Consecutive observations of multi-step forecasts of stock variables, such as integrated realized volatility, can be expected to exhibit relatively persistent short memory dynamics. The empirical autocorrelations of these loss differentials reveal an MA structure with

	Summary statistics					Short memory inference			Long memory inference						
									t_{MAC}			t_{EFB}			
Model	$\overline{z}/\widehat{\sigma}_z$	g_1	g_2	$\widehat{d}_{LW}$	$\widehat{d}_{LPWN}$	t_{DM}	t_{HAC}	t_{FB}	0.7	0.75	0.8	0.2	0.4	0.6	0.8
HAR-RV	0.15	2.03	2.02	**0.23***	**0.20**	**3.07**	**3.25**	**2.77**	1.30	1.46	**1.66**	2.71 (4.23)	3.29 (5.96)	3.29 (8.50)	3.13 (11.34)
HAR-RV-TCJ	0.13	2.03	2.02	**0.19***	0.13	**2.72**	**2.81**	**3.00**	**1.73**	**1.98**	**2.30**	2.98 (3.35)	3.92 (4.87)	3.62 (7.01)	3.59 (9.41)
HAR-RV-TCJ-L	0.10	2.02	2.02	**0.19***	0.10	**2.07** (1.65)	**2.05** (1.65)	**2.86** (2.37)	1.58	**1.85** (1.65)	**2.18**	2.90 (3.35)	3.34 (4.87)	3.21 (7.01)	3.47 (9.41)

Table 11: Predictive ability of the VIX for future RV (evaluated under QLIKE loss). Models excluding the VIX are tested against models including the VIX. See the notes for Table 10.

linearly decaying coefficients. We therefore base all our robust statistics on the LPWN estimator discussed above.[11] Since Chen and Ghysels (2011) and Corsi and Renò (2012) show that the inclusion of leverage effects improves forecasts, we also include a comparison of the HAR-RV-TCJ-L model and the HAR-RV-TCJ-L-VIX model. For details on the HAR-RV-TCJ-L model, see Corsi and Renò (2012) and equation (2) in Bekaert and Hoerova (2014).

Table 10 reports the results on forecast evaluation. Models are estimated using a rolling window of $T_w = 1000$ observations.[12] This implies that the forecast window contains 3883 observations. All DM tests are conducted with one-sided alternatives. In each case, we test that the more complex model outperforms its parsimonious version.

In accordance with our recommendations from Section 5, the t_{EFB} tests are carried out using the Bartlett kernel if MSE loss is considered and with the MQS kernel if QLIKE loss is used. For the sake of a better comparability, we also chose the Bartlett- or the Quadratic spectral kernel for the t_{FB}-statistic, accordingly.

As in the previous literature, the t_{DM}-statistic is implemented using an MA approximation with 44 lags for the forecast horizon of 22 days, cf. for instance Bekaert and Hoerova (2014). For the t_{HAC}-statistic we use an automatic bandwidth selection procedure and the t_{FB}-statistic is computed by using $b = 0.2$ which offers a good trade-off between size control and power, as confirmed in the simulation studies of Sun et al. (2008).

Table 10 reveals that the forecast error loss differentials have long memory with d parameters between 0.14 and 0.24. The results are very similar for the local Whittle and the LPWN estimator. This is a clear indication that memory transmission to the loss differential is taking place, as predicted by Propositions 2 to 5 and the simulations in Section 5.1. Standard DM statistics (t_{DM}, t_{HAC} and t_{FB}) reject the null hypothesis of equal predictive accuracy, thereby confirming the findings in the previous literature. However, if the memory robust statistics in the right panel of Table 10 are taken into account, all evidence for a superior predictive ability of models including the VIX vanishes. Therefore, the previous rejections might be spurious

[11] We choose $R_y = 1$ and $R_w = 0$ concerning the polynomial degrees and a bandwidth $m_d = \lfloor T^{0.8} \rfloor$.

[12] As a robustness check, we repeat the analysis for a larger window of 2500 observations and obtain qualitatively similar results.

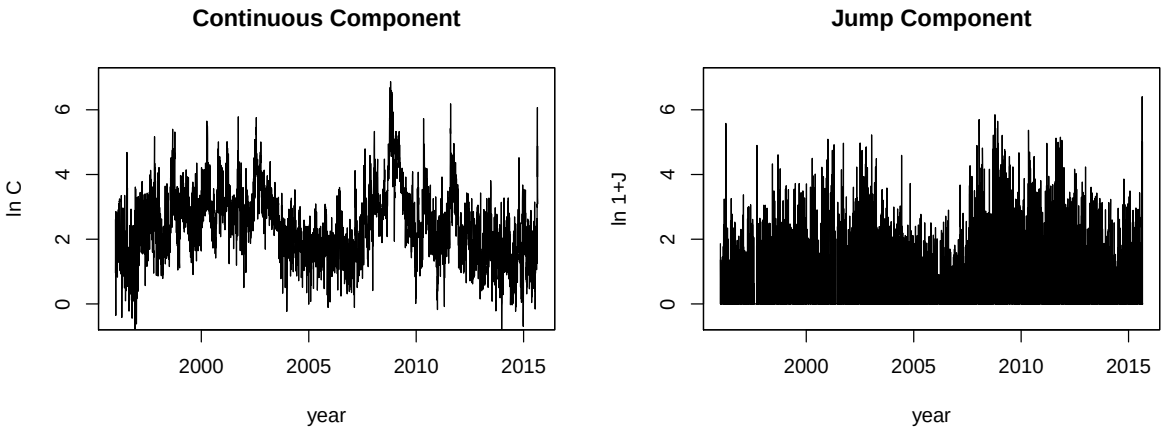

Figure 4: Log continuous component $\ln C_t$ and jump component $\ln(1 + J_t)$ of RV_t.

and reflect the theoretical findings in Proposition 6. In regard of the persistence in the loss differential series the improvements are too small to be considered significant. These findings highlight the importance of long memory robust tests for forecast comparisons in practice.

As a further comparison, we also consider the QLIKE loss function from equation (10) in addition to the MSE. The motivation is that realized volatility is generally considered to be an unbiased, but perturbed proxy of the underlying latent volatility process. It is shown by Patton (2011) that among the commonly employed loss functions only MSE and QLIKE preserve the true ranking of competing forecasts when being evaluated on a perturbed proxy. Even though the propositions presented above only apply for the MSE loss function, the simulations in Section 5.1 clearly show that - given the same setting - memory transmission of a similar magnitude to the MSE case occurs.

Results are reported in Table 11. They suggest that the average standardized forecast error loss differentials are positive and slightly larger in magnitude compared to those for the MSE case in Table 10. Moreover, they have a similar memory structure. From this descriptive viewpoint, results are not sensitive to the choice between the QLIKE and the MSE loss function. When using short memory inference, the null hypothesis of pairwise equal predictive accuracy amongst the models is rejected in all cases. This is also in line with the previous results.

However, when turning to the long memory robust statistics we obtain somewhat different results. On the one hand, the t_{MAC}-statistic rejects for the majority of models and bandwidth parameters. On the other hand, the t_{EFB}-statistic does not generate any rejections. We observe in Table 8 that the tests using QLIKE loss generate considerably more power, and that the power of the t_{EFB}-statistic using the MQS kernel is lower than that of the t_{MAC}-statistic. One could therefore conclude that the inclusion of the VIX improves the accuracy of the forecasts. However, we also find that the t_{MAC}-statistic can be liberal in some situations if the QLIKE loss function is used. Taking these issues into account, the evidence for superior predictive ability of models including the VIX is weak and considerably weaker when considering tests not allowing for the long memory property of the loss differentials.

	Summary statistics					Short memory inference			Long memory inference						
									t_{MAC}			t_{EFB}			
	$\overline{z}/\widehat{\sigma}_z$	g_1	g_2	$\widehat{d}_{LW}$	$\widehat{d}_{LPWN}$	t_{DM}	t_{HAC}	t_{FB}	0.7	0.75	0.8	0.2	0.4	0.6	0.8
$h=1$	0.12	0.41	0.38	**0.09***	0.13	**6.93**	**7.63**	**4.00**	**3.40**	**3.31**	**3.26**	**4.00** (2.61)	**4.07** (3.15)	**4.47** (3.69)	**4.95** (4.23)
$h=5$	0.09	0.26	0.25	**0.07**	0.01	**3.67**	**3.79**	**2.79**	**3.60**	**3.83**	**4.25**	**2.79** (2.05)	**3.98** (2.52)	**5.10** (2.98)	**5.85** (3.39)
$h=22$	0.05	0.29	0.29	**0.36***	**0.34***	0.78 (1.65)	0.91 (1.65)	0.67 (2.09)	0.14	0.15 (1.65)	0.17	0.67 (4.70)	0.93 (5.55)	1.06 (6.41)	1.16 (7.28)

Table 12: Separation of Continuous and Jump Components (evaluated under MSE loss). The forecast performance of the HAR-RV-TCJ model is tested against the HAR-RV for different forecast horizons. Reported are the standardized mean ($\overline{z}/\widehat{\sigma}_z$) and estimated memory parameter ($\widehat{d}$) of the forecast error loss differential. Furthermore, the respective out-of-sample loss of the models (g_1 and g_2) and the results of various DM test statistics are given. Bold-faced values indicate significance at the 5% level and an additional star indicates significance at the 1% level. Critical values of the tests are given in parentheses.

6.2 Separation of Continuous Components and Jump Components

As a second empirical application, we revisit the question whether the HAR-RV-TCJ model from equation (14) leads to a significant improvement in forecast performance compared to the standard HAR-RV-model (13).

The continuous components and jump components - separated using the approach described above - are shown in Figure 4. The occurrence of jumps is often associated with macroeconomic events (cf. Barndorff-Nielsen and Shephard (2006) and Andersen et al. (2007)) and they are observed relatively frequently at about 40% of the days in the sample. The trajectory of the integrated variance follows closely the one of the log-realized volatility series.

Table 12 shows the results of our forecasting exercise for $h \in \{1, 5, 22\}$ steps. Similar to the previous analysis, the t_{DM}-statistic is implemented using an MA approximation including 5, 10 or 44 lags for forecast horizons $h = 1, 5$ and 22, respectively, as is customary in the literature. All other specifications are the same as before. Standard tests (t_{DM}, t_{HAC} and t_{FB}) agree upon rejection of the null hypothesis of equal predictive accuracy in favour of a better performance of the HAR-RV-TCJ model for $h = 1$ and $h = 5$, but not for $h = 22$.

If we consider estimates of the memory parameter, strong (stationary) long memory of 0.34 is only found for $h = 22$. For smaller forecast horizons of $h = 1$ and $h = 5$, LPWN estimates are no longer significantly different from zero since the asymptotic variance is inflated by a multiplicative constant which is also larger for smaller values of d. However, local Whittle estimates remain significant at $\widehat{d}_{LW} = 0.09$ and $\widehat{d}_{LW} = 0.07$ which is qualitatively similar to the results obtained using the LPWN estimator. Therefore, there is again evidence for a transmission of memory to the forecast error loss differential and the rejections of equal predictive accuracy obtained using standard tests might be spurious. Nevertheless, the improvement in forecast

	Summary statistics					Short memory inference			Long memory inference						
									t_{MAC}			t_{EFB}			
	$\overline{z}/\widehat{\sigma}_z$	g_1	g_2	$\widehat{d}_{LW}$	$\widehat{d}_{LPWN}$	t_{DM}	t_{HAC}	t_{FB}	0.7	0.75	0.8	0.2	0.4	0.6	0.8
$h=1$	0.07	1.93	1.93	0.02	0.01	**4.41**	**4.24**	**2.68**	**4.07**	**3.83**	**3.85**	2.60 (2.77)	2.63 (4.10)	3.00 (5.97)	3.76 (7.99)
$h=5$	0.04	1.99	1.99	**0.09***	0.02	1.38	1.42	1.36	1.32	1.38	1.51	1.50 (2.77)	**4.22** (4.10)	3.05 (5.97)	**11.27** (7.99)
$h=22$	−0.01	2.03	2.03	**0.43***	**0.38***	−0.12 (1.65)	−0.14 (1.65)	−0.12 (2.37)	−0.02	−0.02 (1.65)	−0.02	−0.12 (8.85)	−0.21 (11.77)	−0.17 (16.17)	−0.21 (21.60)

Table 13: Separation of Continuous and Jump Components (evaluated under QLIKE loss). See the notes for Table 12.

accuracy is large enough, so that the long memory robust t_{MAC}- and t_{EFB}-statistics reject across the board for $h = 1$ and $h = 5$.

When considering the QLIKE loss function as an alternative to the MSE in Table 13, we find that the memory of the loss differential increases with the forecast horizon, similar to the MSE case. However, the standardized mean of the loss differential is much smaller, which indicates that the improvement achieved by allowing for separate dynamics of the continuous components and the jump components is dependent upon the loss metric. The conventional tests now only reject for $h = 1$. This is confirmed by the t_{MAC}-statistics, but not the t_{EFB} which provides two rejections for the case $h = 5$. Since the estimated memory parameter for $h = 1$ is nearly zero in the QLIKE case, it is likely that the non-rejections of the t_{EFB}-statistic can be attributed to the lower power of the more flexible procedure. Taking these results together, we can therefore confirm that the separation of continuous and jump components indeed improves the forecast performance under MSE loss on daily and weekly horizons. Under QLIKE loss a significant improvement is only found for $h = 1$.

The results obtained so far are based on realized variances calculated including overnight returns (following Bekaert and Hoerova (2014)). Other authors use only intraday returns. As a robustness check, we repeat the analysis excluding overnight returns. In this case, jumps are only detected on about 21 percent of the days. However, the exclusion of overnight returns does not only affect the jump component, but also the RV measure itself. The results of our forecast comparison are shown in Tables 14 to 17 in the appendix. It can be seen that the results are qualitatively similar. If the VIX is included in the HAR-RV models and the forecasts are evaluated using MSE loss, both conventional DM tests and the robust versions reject. Under QLIKE loss we observe that the t_{DM} and t_{HAC}-statistic reject, but not the t_{FB} and the long memory robust versions. This confirms our finding that there is, if at all, weak statistical evidence that the inclusion of the VIX in HAR-RV models improves the predictive accuracy. The memory of the loss differential, however, is found to be even stronger than it is if overnight returns are included. This again highlights the need for memory robust methods.

For the separation of continuous and jump components in the HAR-RV model the results are

also similar to our previous ones. The memory parameters of the loss differentials are a bit lower and the t_{EFB}-statistic under MSE loss rejects only for small bandwidths, whereas the t_{MAC}-statistic now also provides some evidence for predictability with $h = 5$. Overall, we find that our previous finding is robust: the null hypothesis of equal predictive accuracy is rejected for shorter forecast horizons using both conventional and robust tests, but not for $h = 22$.

7 Conclusion

This paper deals with forecast evaluation under long memory. We show that long memory can be transmitted from the forecasts $\widehat{y}_{it}$ and the forecast objective y_t to the forecast error loss differential series z_t in various settings. We demonstrate that the popular test of Diebold and Mariano (1995) is invalidated in these cases. Rejections of the null hypothesis of equal predictive accuracy might therefore be spurious if the series of interest exhibits long memory. For robustification, the MAC estimator of Robinson (2005) and Abadir et al. (2009), as well as the extended fixed-b approach of McElroy and Politis (2012) are discussed.

Simulations verify our theoretical results and demonstrate that the memory transmission extends to other loss functions such as QLIKE, non-Gaussian processes and non-stationary processes. Furthermore, empirical forecast scenarios underline the practical relevance of these issues. Finally, when studying the finite-sample performance of the t_{MAC} and t_{EFB}-statistics, we find that the t_{EFB}-statistic (with the MQS kernel) provides the best size control, whereas the t_{EFB}-statistic using the Bartlett kernel has the best power. Under MSE loss, the size control of all tests is satisfactory in large samples. However, under QLIKE loss we find that only the MQS kernel allows for reliable size control. We therefore recommend to use the Bartlett kernel under MSE loss and in large samples, whereas in smaller samples and under QLIKE loss the MQS kernel should be preferred.

An important example of long memory time series is the realized variance of the S&P 500. It has been the subject of various forecasting exercises. In contrast to previous studies, we only find weak statistical evidence for the hypothesis that the inclusion of the VIX index in HAR-RV-type models leads to an improved forecast performance. Taking the memory of the loss differentials into account reverses the test decisions and suggests that the corresponding findings might be spurious. With regard to the separation of continuous components and jump components, as suggested by Andersen et al. (2007), on the other hand, the improvements in forecast accuracy remain significant at a daily horizon. These examples stress the importance of long memory robust statistics in practice.

Other time series that are routinely found to exhibit long memory include exchange rates, inflation rates, and interest rates (for a recent survey cf. Gil-Alana and Hualde (2009)). The robust test statistics considered here can therefore be helpful in a wider range of applications beyond volatility.

References

Abadir, K. M., Distaso, W., and Giraitis, L. (2009). Two estimators of the long-run variance: Beyond short memory. *Journal of Econometrics*, 150(1):56–70.

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, 89(4):701–720.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96(453):42–55.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858.

Andrews, D. W. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60(4):953–966.

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):253–280.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1):1–37.

Barndorff-Nielsen, O. E. and Shephard, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4(1):1–30.

Becker, R., Clements, A. E., and McClelland, A. (2009). The jump component of S&P 500 volatility and the VIX index. *Journal of Banking & Finance*, 33(6):1033–1038.

Becker, R., Clements, A. E., and White, S. I. (2007). Does implied volatility provide any information beyond that captured in model-based volatility forecasts? *Journal of Banking & Finance*, 31(8):2535–2549.

Bekaert, G. and Hoerova, M. (2014). The VIX, the variance premium and stock market volatility. *Journal of Econometrics*, 183(2):181–192.

Beran, J., Feng, Y., Ghosh, S., and Kulik, R. (2013). *Long memory processes: Probabilistic properties and statistical methods*. Springer London, Limited.

Berkes, I., Rorvath, L., Kokoszka, P., and Shao, Q.-M. (2006). On discriminating between long-range dependence and changes in mean. *The Annals of Statistics*, 34(3):1140–1165.

Bollerslev, T., Osterrieder, D., Sizova, N., and Tauchen, G. (2013). Risk and return: Long-run relations, fractional cointegration, and return predictability. *Journal of Financial Economics*, 108(2):409–424.

Bollerslev, T., Tauchen, G., and Zhou, H. (2009). Expected stock returns and variance risk premia. *Review of Financial Studies*, 22(11):4463–4492.

Busch, T., Christensen, B. J., and Nielsen, M. Ø. (2011). The role of implied volatility in forecasting future realized volatility and jumps in foreign exchange, stock, and bond markets. *Journal of Econometrics*, 160(1):48–57.

Chambers, M. J. (1998). Long memory and aggregation in macroeconomic time series. *International Economic Review*, 39(4):1053–1072.

Chen, X. and Ghysels, E. (2011). News–good or bad–and its impact on volatility predictions over multiple horizons. *Review of Financial Studies*, 24(1):46–81.

Chernov, M. (2007). On the role of risk premia in volatility forecasting. *Journal of Business & Economic Statistics*, 25(4):411–426.

Chiriac, R. and Voev, V. (2011). Modelling and forecasting multivariate realized volatility. *Journal of Applied Econometrics*, 26(6):922–947.

Choi, H.-S. and Kiefer, N. M. (2010). Improving robust model selection tests for dynamic models. *The Econometrics Journal*, 13(2):177–204.

Christensen, B. J. and Nielsen, M. Ø. (2006). Asymptotic normality of narrow-band least squares in the stationary fractional cointegration model and volatility forecasting. *Journal of Econometrics*, 133(1):343–371.

Clark, T. E. (1999). Finite-sample properties of tests for equal forecast accuracy. *Journal of Forecasting*, 18(7):489–504.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

Corsi, F., Pirino, D., and Renò, R. (2010). Threshold bipower variation and the impact of jumps on volatility forecasting. *Journal of Econometrics*, 159(2):276–288.

Corsi, F. and Renò, R. (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *Journal of Business & Economic Statistics*, 30(3):368–380.

Deo, R., Hurvich, C., and Lu, Y. (2006). Forecasting realized volatility using a long-memory stochastic volatility model: estimation, prediction and seasonal adjustment. *Journal of Econometrics*, 131(1):29–58.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics*, 33(1):1–8.

Diebold, F. X. and Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics*, 105:131–159.

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

Dittmann, I. and Granger, C. W. (2002). Properties of nonlinear transformations of fractionally integrated processes. *Journal of Econometrics*, 110(2):113–133.

Fitzsimmons, P. and McElroy, T. (2010). On joint fourier–laplace transforms. *Communications in Statistics – Theory and Methods*, 39(10):1883–1885.

Frederiksen, P., Nielsen, F. S., and Nielsen, M. Ø. (2012). Local polynomial whittle estimation of perturbed fractional processes. *Journal of Econometrics*, 167(2):426–447.

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

Gil-Alana, L. A. and Hualde, J. (2009). Fractional integration and cointegration: an overview and an empirical application. In *Palgrave handbook of econometrics*, pages 434–469. Springer.

Granger, C. W. and Hyung, N. (2004). Occasional structural breaks and long memory with an application to the s&p 500 absolute stock returns. *Journal of Empirical Finance*, 11(3):399–421.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291.

Hou, J. and Perron, P. (2014). Modified local Whittle estimator for long memory processes in the presence of low frequency (and other) contaminations. *Journal of Econometrics*, 182(2):309–328.

Hualde, J. and Velasco, C. (2008). Distribution-free tests of fractional cointegration. *Econometric Theory*, 24(01):216–255.

Kechagias, S. and Pipiras, V. (2015). Definitions and representations of multivariate long-range dependent time series. *Journal of Time Series Analysis*, 36(1):1–25.

Kiefer, N. M. and Vogelsang, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, 21(6):1130–1164.

Kruse, R. (2015). A modified test against spurious long memory. *Economics Letters*, 135:34–38.

Leschinski, C. (2017). On the memory of products of long range dependent time series. *Economics Letters*, 153:72–76.

Li, J. and Patton, A. J. (2015). Asymptotic inference about predictive accuracy using high frequency data. *unpublished*.

Lu, Y. K. and Perron, P. (2010). Modeling and forecasting stock return volatility using a random level shift model. *Journal of Empirical Finance*, 17(1):138–156.

Mariano, R. S. and Preve, D. (2012). Statistical tests for multiple forecast comparison. *Journal of Econometrics*, 169(1):123–130.

Martens, M., Van Dijk, D., and De Pooter, M. (2009). Forecasting S&P 500 volatility: Long memory, level shifts, leverage effects, day-of-the-week seasonality, and macroeconomic announcements. *International Journal of Forecasting*, 25(2):282–303.

McElroy, T. and Politis, D. N. (2012). Fixed-b asymptotics for the studentized mean from time series with short, long, or negative memory. *Econometric Theory*, 28(2):471–481.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Nielsen, M. Ø. (2007). Local Whittle analysis of stationary fractional cointegration and the implied–realized volatility relation. *Journal of Business & Economic Statistics*, 25(4):427–446.

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

Patton, A. J. and Sheppard, K. (2009). Evaluating volatility and correlation forecasts. In *Handbook of financial time series*, pages 801–838. Springer.

Perron, P. and Qu, Z. (2010). Long-memory and level shifts in the volatility of stock market return indices. *Journal of Business & Economic Statistics*, 28(2):275–290.

Phillips, P. C. B. and Kim, C. S. (2007). Long-run covariance matrices for fractionally integrated processes. *Econometric Theory*, 23(6):1233–1247.

Qu, Z. (2011). A test against spurious long memory. *Journal of Business & Economic Statistics*, 29(3):423–438.

Robinson, P. M. (2005). Robust covariance matrix estimation: HAC estimates with long memory/antipersistence correction. *Econometric Theory*, 21(1):171–180.

Rossi, B. (2005). Testing long-horizon predictive ability with high persistence, and the meese–rogoff puzzle*. *International Economic Review*, 46(1):61–92.

Sun, Y., Phillips, P. C., and Jin, S. (2008). Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing. *Econometrica*, 76(1):175–194.

Varneskov, R. T. and Perron, P. (2017). Combining long memory and level shifts in modelling and forecasting the volatility of asset returns. *Quantitative Finance*, pages 1–23.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5):1067–1084.

Appendix

Proofs

Proof (Proposition 2). *By defining $a_t^* = a_t - \mu_a$, for $a_t \in \{y_t, \widehat{y}_{1t}, \widehat{y}_{2t}\}$, the loss differential z_t in (7) can be re-expressed as*

$$\begin{aligned} z_t =& -2y_t(\widehat{y}_{1t} - \widehat{y}_{2t}) + \widehat{y}_{1t}^2 - \widehat{y}_{2t}^2 \\ =& -2(y_t^* + \mu_y)[\widehat{y}_{1t}^* + \mu_1 - \widehat{y}_{2t}^* - \mu_2] + (\widehat{y}_{1t}^* + \mu_1)^2 - (\widehat{y}_{2t}^* + \mu_2)^2 \\ =& -2\{y_t^*\widehat{y}_{1t}^* + \mu_1 y_t^* - y_t^*\widehat{y}_{2t}^* - y_t^*\mu_2 + \mu_y\widehat{y}_{1t}^* + \mu_y\mu_1 - \widehat{y}_{2t}^*\mu_y - \mu_2\mu_y\} \\ & + \widehat{y}_{1t}^{*^2} + 2\widehat{y}_{1t}^*\mu_1 + \mu_1^2 - \widehat{y}_{2t}^{*^2} - 2\widehat{y}_{2t}^*\mu_2 - \mu_2^2 \\ =& \underbrace{-2[y_t^*(\mu_1 - \mu_2) + \widehat{y}_{1t}^*(\mu_y - \mu_1) - \widehat{y}_{2t}^*(\mu_y - \mu_2)]}_{I} - \underbrace{2[y_t^*(\widehat{y}_{1t}^* - \widehat{y}_{2t}^*)]}_{II} + \underbrace{\widehat{y}_{1t}^{*^2} - \widehat{y}_{2t}^{*^2}}_{III} + const. \end{aligned} \tag{15}$$

Proposition 3 in Chambers (1998) states that the memory of a linear combination of fractionally integrated processes is equal to the maximum of the memory orders of the components. As discussed in Leschinski (2017), this result also applies for long memory processes in general since the proof is only based on the long memory properties of the fractionally integrated processes. We can therefore also apply it to (15). In order to determine the memory of the forecast error loss differential z_t, we have to determine the memory orders of the three individual components I, II and III in the linear combination.

Regarding I, we have $y_t^* \sim LM(d_y)$, $\widehat{y}_{1t}^* \sim LM(d_1)$ and $\widehat{y}_{2t}^* \sim LM(d_2)$. For terms II and III, we refer to Proposition 1 from Leschinski (2017). We thus have for $i \in \{1, 2\}$

$$y_t^*\widehat{y}_{it}^* \sim \begin{cases} LM(\max\{d_y + d_i - 1/2, 0\}), & \text{if } S_{y,\widehat{y}_i} \neq 0 \\ LM(d_y + d_i - 1/2), & \text{if } S_{y,\widehat{y}_i} = 0 \end{cases} \tag{16}$$

$$\text{and} \qquad \widehat{y}_{it}^{*^2} \sim LM(\max\{2d_i - 1/2, 0\}). \tag{17}$$

Further note that

$$d_y > d_y + d_i - 1/2 \;\; \text{and} \;\; d_i > d_y + d_i - 1/2 \tag{18}$$

and

$$d_i > 2d_i - 1/2, \tag{19}$$

since $0 \leq d_a < 1/2$ for $a \in \{y, 1, 2\}$.

Using these properties, we can determine the memory d_z in (15) via a case-by-case analysis.

1. *First, if $\mu_1 \neq \mu_2 \neq \mu_y$ the memory of the original terms dominates because of (18) and (19) and we obtain $d_z = \max\{d_y,\ d_1,\ d_2\}$.*

2. *Second, if $\mu_1 = \mu_2 \neq \mu_y$, then y_t^* drops out from (15), but the two forecasts $\widehat{y}_{1t}$ and $\widehat{y}_{2t}$ remain. From (18) and (19), we have that d_1 and d_2 dominate their transformations*

leading to the result $d_z = \max\{d_1,\ d_2\}$.

3. *Third, if* $\mu_1 = \mu_y \neq \mu_2$, *the forecast* $\widehat{y}^*_{1t}$ *vanishes and* d_2 *and* d_y *dominate their reduced counterparts by (18) and (19), so that* $d_z = \max\{2d_1 - 1/2,\ d_2,\ d_y\}$.

4. *Fourth, by the same arguments just as before,* $d_z = \max\{2d_2 - 1/2,\ d_1,\ d_y\}$ *if* $\mu_2 = \mu_y \neq \mu_1$.

5. *Finally, if* $\mu_1 = \mu_2 = \mu_y$, *the forecast objective* y^*_t *as well as both forecasts* $\widehat{y}^*_{1t}$ *and* $\widehat{y}^*_{2t}$ *drop from (15). The memory of the loss differential is therefore the maximum of the memory orders in the remaining four terms in* II *and* III *that are given in (16) and (17). Furthermore, the memory of the squared series given in (17) is always non-negative from Corollary 1 in Leschinski (2017) and a linear combination of an antipersistent process with an* $LM(0)$ *series is* $LM(0)$, *from Proposition 3 of Chambers (1998). Therefore, the lower bound for* d_z *is zero and*

$$d_z = \max\{2\max\{d_1,\ d_2\} - 1/2,\ d_y + \max\{d_1,\ d_2\} - 1/2,\ 0\}\,. \ \square$$

Proof (Proposition 3)**.** *For the case that common long memory is permitted, we consider three possible situations: CLM between the forecasts $\widehat{y}_{1t}$ and $\widehat{y}_{2t}$, CLM between the forecast objective y_t and one of the forecasts $\widehat{y}_{1t}$ or $\widehat{y}_{2t}$ and finally CLM between y_t and each $\widehat{y}_{1t}$ and $\widehat{y}_{2t}$.*

First, note that as a direct consequence of Assumption 3, we have

$$\mu_i = \beta_i + \xi_i \mu_x \tag{20}$$

and

$$\mu_y = \beta_y + \xi_y \mu_x. \tag{21}$$

We can now re-express the forecast error loss differential z_t in (15) for each possible CLM relationship. In all cases, tedious algebraic steps are not reported to save space.

1. *In the case of CLM between $\widehat{y}_{1t}$ and $\widehat{y}_{2t}$, we have*

$$\begin{aligned} z_t = &-2\{y_t^*(\mu_1-\mu_2) + x_t^*[\xi_1(\mu_y-\mu_1) - \xi_2(\mu_y-\mu_2)] + x_t^* y_t^*(\xi_1-\xi_2) - x_t^*(\xi_1\varepsilon_{1t} - \xi_2\varepsilon_{2t}) \\ &+ \varepsilon_{1t}(\mu_y-\mu_1) - \varepsilon_{2t}(\mu_y-\mu_2) + \mu_x(\varepsilon_{1t}\xi_1 - \varepsilon_{2t}\xi_2) + y_t^*(\varepsilon_{1t}-\varepsilon_{2t})\} \\ &+ x_t^{*^2}(\xi_1^2-\xi_2^2) + \varepsilon_{1t}^2 - \varepsilon_{2t}^2 + 2\mu_x(\varepsilon_{1t}\xi_1 - \varepsilon_{2t}\xi_2) + const. \end{aligned} \tag{22}$$

2. *If the forecast objective y_t and one of the $\widehat{y}_{it}$ have CLM, we have for $\widehat{y}_{1t}$:*

$$\begin{aligned} z_t = &-2\{x_t^*[(\mu_y-\mu_1)\xi_1 + \xi_y(\mu_1-\mu_2)] - \widehat{y}_{2t}^*[\mu_y-\mu_2] - \xi_y x_t^* \widehat{y}_{2t}^* + x_t^*[\varepsilon_{1t}(\xi_y-\xi_1) + \xi_1\eta_t] \\ &+ \varepsilon_{1t}(\xi_y\mu_x - \mu_1) + \eta_t(\mu_1-\mu_2) + \varepsilon_{1t}\eta_t - \widehat{y}_{2t}^*\eta_t\} \\ &- (2\xi_1\xi_y - \xi_1^2)x_t^{*^2} + \varepsilon_{1t}^2 - \widehat{y}_{2t}^{*^2} - 2\beta_y\varepsilon_{1t} + const. \end{aligned} \tag{23}$$

   The result for CLM between y_t and $\widehat{y}_{2t}$ is entirely analogous, but with index "1" being replaced by "2".

3. *Finally, if y_t has CLM with both $\widehat{y}_{1t}$ and $\widehat{y}_{2t}$, we have:*

$$\begin{aligned} z_t = &-2\big\{x_t^*[\xi_1(\mu_y-\mu_1) - \xi_2(\mu_y-\mu_2) + \xi_y(\mu_1-\mu_2)] \\ &+ x_t^*[(\xi_y-\xi_1)\varepsilon_{1t} - (\xi_y-\xi_2)\varepsilon_{2t} + (\xi_1-\xi_2)\eta_t] \\ &+ x_t^{*2}[\xi_y(\xi_1-\xi_2) - \frac{1}{2}(\xi_1^2-\xi_2^2)] \\ &+ \varepsilon_{1t}(\mu_y-\mu_1) - \varepsilon_{2t}(\mu_y+\mu_2) + \mu_x(\xi_1\varepsilon_{1t} + \xi_2\varepsilon_{2t}) + \eta_t(\varepsilon_{1t}-\varepsilon_{2t}) + \eta_t[\mu_1-\mu_2]\big\} \\ &+ \varepsilon_{1t}^2 - \varepsilon_{2t}^2 + 2\mu_x(\xi_1\varepsilon_{1t} - \xi_2\varepsilon_{2t}) + const. \end{aligned} \tag{24}$$

As in the proof of Proposition 2, we can now determine the memory orders of z_t in (22), (23) and (24) by first considering the memory of each term in each of the linear combinations and

then by applying Proposition 3 of Chambers (1998) thereafter. However, note that

$$y_t^*(\mu_1 - \mu_2) + x_t^*[\xi_1(\mu_y - \mu_1) - \xi_2(\mu_y - \mu_2)] \;\; \textit{in (22)},$$

$$x_t^*[(\mu_y - \mu_1)\xi_1 + \xi_y(\mu_1 - \mu_2)] - \widehat{y}_{2t}^*(\mu_y - \mu_2) \;\; \textit{in (23)}$$

and

$$x_t^*[\xi_1(\mu_y - \mu_1) - \xi_2(\mu_y - \mu_2) + \xi_y(\mu_1 - \mu_2)] \;\; \textit{in (24)}$$

have the same structure as

$$y_t^*(\mu_1 - \mu_2) + \widehat{y}_{1t}^*(\mu_y - \mu_1) - \widehat{y}_{2t}^*(\mu_y - \mu_2) \;\; \textit{in (15)}$$

and that all of the other non-constant terms in (22), (23) and (24) are either squares or products of demeaned series, so that their memory is reduced according to Proposition 1 from Leschinski (2017). From Assumption 3, x_t^* is the common factor driving the series with CLM and from $d_x > d_{\varepsilon_1}, d_{\varepsilon_2}, d_\eta$ and the dominance of the largest memory in a linear combination from Proposition 3 in Chambers (1998), x_t^* has the same memory as the series involved in the CLM relationship. Now from (18) and (19), the reduced memory of the product series and the squared series is dominated by that of either x_t^*, y_t^*, $\widehat{y}_{1t}^*$ or $\widehat{y}_{2t}^*$. Therefore, whenever a bias term is non-zero, the memory of the linear combination can be no smaller than that of the respective original series.

To obtain the results in Proposition 3, set the terms in square brackets in equations (22), (23) and (15) equal to zero and solve for the quotient of the factor loadings. This determines the transmission of the memory of x_t^*. For the effect of the series that is not involved in the CLM relationship, we impose the restrictions $\mu_1 \neq \mu_2$ and $\mu_i \neq \mu_y$, as stated in the proposition. □

Proof (Proposition 4)**.** *The results in Proposition 4 follow directly from equations (22), (23) and (24), above. For (22) the terms in square brackets can be re-expressed as*

$$[\xi_1(\mu_y - \mu_1) - \xi_2(\mu_y - \mu_2)] = (\xi_1 - \xi_2)\mu_y + \xi_2\mu_2 - \xi_1\mu_1.$$

Obviously, for $\xi_1 = \xi_2$, this is reduced to $\xi_2\mu_2 - \xi_1\mu_1$, which does not vanish since $\mu_1 \neq \mu_2$. The other cases are treated entirely analogous. For (23) we have

$$[(\mu_y - \mu_1)\xi_1 + \xi_y(\mu_1 - \mu_2)] = \xi_1\mu_y - \xi_y\mu_2,$$

and in (24)

$$[\xi_1(\mu_y - \mu_1) - \xi_2(\mu_y - \mu_2) + \xi_y(\mu_1 - \mu_2)] = (\xi_1 - \xi_2)\mu_y - (\xi_1 - \xi_y)\mu_1 + (\xi_2 - \xi_y)\mu_2 = 0,$$

so that x_t^* drops out and the memory is reduced. □

Proof (Proposition 5)**.** *Under the assumptions of Proposition 3, (22) is reduced to*

$$\begin{aligned} z_t &= -2\{-x_t^*(\xi_1\varepsilon_{1t} - \xi_2\varepsilon_{2t}) + y_t^*(\varepsilon_{1t} - \varepsilon_{2t})\} + \varepsilon_{1t}^2 - \varepsilon_{2t}^2 + const, \\ &= -2\{-\underbrace{\xi_1 x_t^*\varepsilon_{1t}}_{I} + \underbrace{\xi_2 x_t^*\varepsilon_{2t}}_{II} + \underbrace{y_t^*\varepsilon_{1t}}_{III} - \underbrace{y_t^*\varepsilon_{2t}}_{IV}\} + \underbrace{\varepsilon_{1t}^2}_{V} - \underbrace{\varepsilon_{2t}^2}_{VI} + const, \end{aligned} \tag{25}$$

(23) becomes

$$\begin{aligned} z_t &= -2\{-x_t^*(\xi_y\widehat{y}_{2t}^* - \xi_1\eta_t) + (\varepsilon_{1t} - \widehat{y}_{2t}^*)\eta_t + \varepsilon_{1t}(\xi_y\mu_x - \mu_1)\} + \varepsilon_{1t}^2 - \widehat{y}_{2t}^{*^2} - 2\beta_y\varepsilon_{1t} - \xi_1\xi_y x_t^{*^2} + const, \\ &= -2\{-\underbrace{\xi_y x_t^*\widehat{y}_{2t}^*}_{I} + \underbrace{\xi_1 x_t^*\eta_t}_{II} + \underbrace{\varepsilon_{1t}\eta_t}_{III} - \underbrace{\widehat{y}_{2t}^*\eta_t}_{IV} + \underbrace{\varepsilon_{1t}(\xi_y\mu_x - \mu_1)}_{V}\} + \underbrace{\varepsilon_{1t}^2}_{VI} - \underbrace{\widehat{y}_{2t}^{*^2}}_{VII} - \underbrace{2\beta_y\varepsilon_{1t}}_{VIII} - \underbrace{\xi_1\xi_y x_t^{*^2}}_{IX} + const, \end{aligned} \tag{26}$$

and finally (24) is

$$\begin{aligned} z_t &= -2(\varepsilon_{1t} - \varepsilon_{2t})\eta_t + \varepsilon_{1t}^2 - \varepsilon_{2t}^2 + const, \\ &= -2\underbrace{\varepsilon_{1t}\eta_t}_{I} + \underbrace{2\varepsilon_{2t}\eta_t}_{II} + \underbrace{\varepsilon_{1t}^2}_{III} - \underbrace{\varepsilon_{2t}^2}_{IV} + const. \end{aligned} \tag{27}$$

We can now proceed as in the proof of Proposition 2 and infer the memory orders of each term in the respective linear combination from Proposition 1 and then determine the maximum as in Proposition 3 in Chambers (1998).

In the following, we label the terms appearing in each of the equations by consecutive letters with the equation number as an index. For the terms in (25), we have

$$I_{25} \sim \begin{cases} LM(\max\{d_x + d_{\varepsilon_1} - 1/2,\ 0\}), & \text{if } S_{x,\varepsilon_1} \neq 0 \\ LM(d_x + d_{\varepsilon_1} - 1/2), & \text{if } S_{x,\varepsilon_1} = 0 \end{cases}$$

$$II_{25} \sim \begin{cases} LM(\max\{d_x + d_{\varepsilon_2} - 1/2,\ 0\}), & \text{if } S_{x,\varepsilon_2} \neq 0 \\ LM(d_x + d_{\varepsilon_2} - 1/2), & \text{if } S_{x,\varepsilon_2} = 0 \end{cases}$$

$$III_{25} \sim \begin{cases} LM(\max\{d_y + d_{\varepsilon_1} - 1/2,\ 0\}), & \text{if } S_{y,\varepsilon_1} \neq 0 \\ LM(d_y + d_{\varepsilon_1} - 1/2), & \text{if } S_{y,\varepsilon_1} = 0 \end{cases}$$

$$IV_{25} \sim \begin{cases} LM(\max\{d_y + d_{\varepsilon_2} - 1/2,\ 0\}), & \text{if } S_{y,\varepsilon_2} \neq 0 \\ LM(d_y + d_{\varepsilon_2} - 1/2), & \text{if } S_{y,\varepsilon_2} = 0 \end{cases}$$

$$V_{25} \sim LM(\max\{2d_{\varepsilon_1} - 1/2,\ 0\})$$

$$\text{and} \qquad VI_{25} \sim LM(\max\{2d_{\varepsilon_2} - 1/2,\ 0\}).$$

Since by definition $d_x > d_{\varepsilon_i}$, the memory of V_{25} and VI_{25} is always of a lower order than that of I_{25} and II_{25}. As in the proof of Proposition 2, the squares in terms V_{25} and VI_{25} establish zero as the lower bound of d_z. Therefore, we have

$$d_z = \max\{\max\{d_x,\ d_y\} + \max\{d_{\varepsilon_1},\ d_{\varepsilon_2}\} - 1/2,\ 0\}.$$

Similarly, in (26), we have

$$I_{26} \sim \begin{cases} LM(\max\{d_x + d_2 - 1/2,\ 0\}), & \text{if } S_{x,\widehat{y}_2} \neq 0 \\ LM(d_x + d_2 - 1/2), & \text{if } S_{x,\widehat{y}_2} = 0 \end{cases}$$

$$II_{26} \sim \begin{cases} LM(\max\{d_x + d_\eta - 1/2,\ 0\}), & \text{if } S_{x,\eta} \neq 0 \\ LM(d_x + d_\eta - 1/2), & \text{if } S_{x,\eta} = 0 \end{cases}$$

$$III_{26} \sim \begin{cases} LM(\max\{d_{\varepsilon_1} + d_\eta - 1/2,\ 0\}), & \text{if } S_{\varepsilon_1,\eta} \neq 0 \\ LM(d_{\varepsilon_1} + d_\eta - 1/2), & \text{if } S_{\varepsilon_1,\eta} = 0 \end{cases}$$

$$IV_{26} \sim \begin{cases} LM(\max\{d_2 + d_\eta - 1/2,\ 0\}), & \text{if } S_{\widehat{y}_2,\eta} \neq 0 \\ LM(d_2 + d_\eta - 1/2), & \text{if } S_{\widehat{y}_2,\eta} = 0 \end{cases}$$

$$V_{26} \sim LM(d_{\varepsilon_1})$$

$$VI_{26} \sim LM(\max\{2d_{\varepsilon_1} - 1/2,\ 0\})$$

$$VII_{26} \sim LM(\max\{2d_2 - 1/2,\ 0\})$$

$$VIII_{26} \sim LM(d_{\varepsilon_1})$$

$$\text{and} \qquad IX_{26} \sim LM(\max\{2d_x - 1/2, 0\}).$$

Here, V_{26} can be disregarded since it is of the same order as $VIII_{26}$. $VIII_{26}$ dominates VI_{26}, because $d_{\varepsilon_1} < 1/2$. Finally, as $d_{\varepsilon_1} < d_x$ holds by assumption, III_{26} is dominated by II_{26} and $d_\eta < d_x$, so that IX_{26} dominates II_{26}. Therefore,

$$d_z = \max\{d_2 + \max\{d_x, d_\eta\} - 1/2,\ 2\max\{d_x, d_2\} - 1/2,\ d_{\varepsilon_1}\}.$$

As before, for the case of CLM between y_t and $\widehat{y}_{2t}$, the proof is entirely analogous, but with index "1" replaced by "2" and vice versa.

Finally, in (27), we have

$$I_{27} \sim \begin{cases} LM(\max\{d_\eta + d_{\varepsilon_1} - 1/2,\ 0\}), & \text{if } S_{\eta,\varepsilon_1} \neq 0 \\ LM(d_\eta + d_{\varepsilon_1} - 1/2), & \text{if } S_{\eta,\varepsilon_1} = 0 \end{cases}$$

$$II_{27} \sim \begin{cases} LM(\max\{d_\eta + d_{\varepsilon_2} - 1/2,\ 0\}), & \text{if } S_{\eta,\varepsilon_1} \neq 0 \\ LM(d_\eta + d_{\varepsilon_2} - 1/2), & \text{if } S_{\eta,\varepsilon_2} = 0 \end{cases}$$

$$III_{27} \sim LM(\max\{2d_{\varepsilon_1} - 1/2,\ 0\})$$

$$IV_{27} \sim LM(\max\{2d_{\varepsilon_2} - 1/2,\ 0\}).$$

Here, no further simplifications can be made since we do not impose restrictions on the relationship between d_η, d_{ε_1} and d_{ε_2}, so that

$$d_z = \max\{d_\eta + \max\{d_{\varepsilon_1},\ d_{\varepsilon_2}\} - 1/2,\ 2\max\{d_{\varepsilon_1},\ d_{\varepsilon_2}\} - 1/2,\ 0\},$$

where again the zero is established as the lower bound by the squares in III_{27} and IV_{27}. □

Proof (Proposition 6). *Under short memory, the t_{HAC}-statistic is given by*

$$t_{HAC} = T^{1/2}\frac{\bar{z}}{\sqrt{\widehat{V}_{HAC}}},$$

with $\widehat{V}_{HAC} = \sum_{j=-T+1}^{T-1} k\left(\frac{j}{B}\right)\widehat{\gamma}_z(j)$ and B being the bandwidth satisfying $B \to \infty$ and $B = O(T^{1-\epsilon})$ for some $\epsilon > 0$. From Abadir et al. (2009), the appropriately scaled long-run variance estimator for a long memory processes is given by $B^{-1-2d}\sum_{i,j=1}^{B}\widehat{\gamma}_z(|i-j|)$, see equation (2.2) in Abadir et al. (2009). Corresponding long memory robust HAC-type estimators (with a Bartlett kernel, for instance) take the form

$$\widehat{V}_{HAC,d} = B^{-2d}\left(\widehat{\gamma}_z(0) + 2\sum_{j=1}^{B}(1-j/B)\widehat{\gamma}_z(j)\right).$$

The long memory robust $t_{HAC,d}$-statistic is then given by

$$t_{HAC,d} = T^{1/2-d}\frac{\bar{z}}{\sqrt{\widehat{V}_{HAC,d}}}.$$

We can therefore write

$$t_{HAC,d} = T^{1/2}T^{-d}\frac{\bar{z}}{\sqrt{B^{-2d}\widehat{V}_{HAC}}} = \frac{T^{-d}}{B^{-d}}t_{HAC}$$

and thus,

$$t_{HAC} = \frac{T^d}{B^d}t_{HAC,d}.$$

The short memory t_{HAC}-statistic is inflated by the scaling factor $T^d/B^d = O(T^{d\epsilon})$. This leads directly to the divergence of the $_{HAC}$-statistic ($t_{HAC} \to \infty$ as $T \to \infty$) which implies that

$$\lim_{T\to\infty} P(|t_{HAC}| > c_{1-\alpha/2,d}) = 1$$

for all values of $d \in (0, 1/4) \cup (1/4, 1/2)$. For $0 < d < 1/4$, $c_{1-\alpha/2,d}$ is the critical value from the $N(0,1)$-distribution, while for $1/4 < d < 1/2$, the critical value (depending with d) stems from the well-defined Rosenblatt distribution, see Abadir et al. (2009). The proof is analogous for other kernels and thus omitted. □

Additional Material for the Empirical Applications

		Summary statistics				Short memory inference			Long memory inference						
									t_{MAC}			t_{EFB}			
Model	$\overline{z}/\widehat{\sigma}_z$	g_1	g_2	$\widehat{d}_{LW}$	$\widehat{d}_{LPWN}$	t_{DM}	t_{HAC}	t_{FB}	0.7	0.75	0.8	0.2	0.4	0.6	0.8
HAR-RV	0.08	0.28	0.27	**0.36***	**0.34***	1.23	1.36	1.08	0.25	0.28	0.33	1.08	1.40	2.10	2.33
												(4.70)	(5.55)	(6.41)	(7.28)
HAR-RV-TCJ	0.07	0.28	0.27	**0.36***	**0.31***	1.10	1.22	0.97	0.27	0.31	0.36	0.97	1.35	2.04	2.02
												(4.70)	(5.55)	(6.41)	(7.28)
HAR-RV-TCJ-L	0.05	0.27	0.26	**0.30***	**0.27***	0.79	0.87	0.78	0.24	0.27	0.32	0.78	1.10	1.61	1.56
						(1.65)	(1.65)	(2.09)		(1.65)		(4.70)	(5.55)	(6.41)	(7.28)

Table 14: Predictive ability of models including the VIX for future RV calculated excluding overnight returns (evaluated under MSE loss). See the notes for Table 10.

		Summary statistics				Short memory inference			Long memory inference						
									t_{MAC}			t_{EFB}			
Model	$\overline{z}/\widehat{\sigma}_z$	g_1	g_2	$\widehat{d}_{LW}$	$\widehat{d}_{LPWN}$	t_{DM}	t_{HAC}	t_{FB}	0.7	0.75	0.8	0.2	0.4	0.6	0.8
HAR-RV	0.13	1.92	1.92	**0.30***	**0.21**	**2.27**	**2.40**	1.93	1.00	1.15	1.32	2.00	4.91	6.99	3.66
												(4.23)	(5.96)	(8.50)	(11.34)
HAR-RV-TCJ	0.12	1.92	1.92	**0.25***	**0.23***	**2.22**	**2.40**	1.76	0.85	0.94	1.06	1.82	3.95	3.34	2.83
												(4.23)	(5.96)	(8.50)	(11.34)
HAR-RV-TCJ-L	0.10	1.92	1.92	**0.25***	**0.20**	**1.92**	**2.01**	1.72	0.85	0.97	1.11	1.79	3.47	2.95	2.68
						(1.65)	(1.65)	(2.37)		(1.65)		(4.23)	(5.96)	(8.50)	(11.34)

Table 15: Predictive ability of models including the VIX for future RV calculated excluding overnight returns (evaluated under QLIKE loss). See the notes for Table 11.

	Summary statistics					Short memory inference			Long memory inference						
									t_{MAC}			t_{EFB}			
	$\overline{z}/\widehat{\sigma}_z$	g_1	g_2	$\widehat{d}_{LW}$	$\widehat{d}_{LPWN}$	t_{DM}	t_{HAC}	t_{FB}	0.7	0.75	0.8	0.2	0.4	0.6	0.8
$h=1$	0.07	0.31	0.30	**0.08**	0.11	**4.16**	**4.59**	**2.79**	**2.27**	**2.30**	**2.35**	**2.79** (2.61)	2.54 (3.15)	2.66 (3.69)	2.96 (4.23)
$h=5$	0.069	0.229	0.222	0.057	0.000	**2.368**	**2.398**	**2.134**	**2.474**	**2.704**	**3.155**	**2.134** (2.05)	2.267 (2.52)	2.671 (2.98)	3.093 (3.39)
$h=22$	0.05	0.28	0.28	**0.23***	0.16	1.19 (1.65)	1.18 (1.65)	1.48 (3.40)	0.55	0.63 (1.65)	0.75	1.48 (3.40)	1.64 (4.06)	1.96 (4.75)	2.12 (5.39)

Table 16: Separation of continuous and jump components for RV calculated excluding overnight returns (evaluated under MSE loss). See the notes for Table 12.

	Summary statistics					Short memory inference			Long memory inference						
									t_{MAC}			t_{EFB}			
	$\overline{z}/\widehat{\sigma}_z$	g_1	g_2	$\widehat{d}_{LW}$	$\widehat{d}_{LPWN}$	t_{DM}	t_{HAC}	t_{FB}	0.7	0.75	0.8	0.2	0.4	0.6	0.8
$h=1$	0.03	1.82	1.82	0.04	0.05	**2.38**	**2.24**	1.95	**1.76**	**1.68**	1.62	1.97 (2.77)	2.83 (4.10)	5.80 (5.97)	6.08 (7.99)
$h=5$	0.04	1.88	1.88	0.05	0.01	**1.77**	**1.79**	1.92	**1.69**	**1.72**	**1.88**	2.10 (2.77)	2.51 (4.10)	4.65 (5.97)	4.41 (7.99)
$h=22$	0.06	1.92	1.92	**0.22***	0.16	1.53 (1.65)	1.39 (1.65)	4.32 (2.37)	0.66	0.75 (1.65)	0.86	**26.03** (4.23)	3.68 (5.96)	5.36 (8.50)	10.29 (11.34)

Table 17: Separation of continuous and jump components for RV calculated excluding overnight returns (evaluated under QLIKE loss). See the notes for Table 13.