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

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

[a]Rijksuniversiteit Groningen [b]CREATES, Aarhus University and [c]Leibniz University Hannover

November 29, 2016

Abstract

This paper extends the popular Diebold-Mariano test to situations when the forecast error loss differential exhibits long memory. It is shown that 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 mainly depends on the (un)biasedness of the forecasts and whether the involved series share common long memory. Further 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 Monte Carlo study provides a novel comparison of these robust statistics. As empirical applications, we conduct forecast comparison tests for the realized volatility of the Standard and Poors 500 index among recent extensions of the heterogeneous autoregressive model. While we find that forecasts improve significantly if jumps in the log-price process are considered separately from continuous components, improvements achieved by the inclusion of implied volatility turn out to be insignificant in most situations.

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

JEL classification: C22; C52; C53

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

*The online appendix to this paper is available here. We would like to thank Philipp Sibbertsen, Karim Abadir, Guillaume Chevillion, Mauro Costantini, Matei Demetrescu, Niels Haldrup, Uwe Hassler, Tucker McElroy, Andrew Patton and Uta Pigorsch 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.

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 of the forecaster be given by $g(y_t, \widehat{y}_{it}) \geq 0$ for $i = 1, 2$. The forecast error loss differential is then denoted by $z_t = g(y_t, \widehat{y}_{1t}) - g(y_t, \widehat{y}_{2t})$.

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)). Another extension of the DM test is proposed by Rossi (2005), who develops a DM test under near unit root asymptotics. However, all of these extensions fall into the classical $I(0)/I(1)$ framework.

In this paper, we study the situation if these assumptions on the loss differential do not apply and instead z_t follows a long memory process. 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 consider the case of a mean squared error (MSE) loss function and give conditions under which the transmission occurs and characterize the memory properties of the forecast error loss differential. As a second contribution, we show (both theoretically and via simulations) that the original DM test is invalidated in this case and suffers from severe upward size distortions. Third, we study two simple extensions of the DM statistic that allow valid inference under long (and short) memory. These extensions are 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. Since these tests build on a restriction on the mean, the results allow broader conclusions for similar inference problems (besides the Diebold-Mariano test) which is an interesting topic in its own right. We compare several bandwidth and kernel choices that allow recommendations for practical applications.

Our fourth contribution is an empirical application where 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 ability 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.

The second issue we tackle relates to earlier work by Andersen et al. (2007) and Corsi et al. (2010), among others, 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

$$z_t = \mu_z + \sum_{j=0}^{\infty} \theta_j v_{t-j} \ , \tag{1}$$

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

$$t_{DM} = T^{\delta} \frac{\bar{z}}{\sqrt{\widehat{V}}} \ . \tag{2}$$

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 estimators for the long-run variance. In particular,

$$\widehat{V}_{HAC} = \sum_{j=-T+1}^{T-1} k\left(\frac{j}{B}\right) \widehat{\gamma}_z(j) \,, \tag{3}$$

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

$$t_{HAC} = T^{1/2} \frac{\bar{z}}{\sqrt{\widehat{V}_{HAC}}}. \tag{4}$$

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.

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

as $\sqrt{V}$ in (2) cancels out as long as $\widehat{V} \xrightarrow{p} V$ holds. 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

$$t_{FB} = T^{1/2} \frac{\bar{z}}{\sqrt{\widehat{V}(k,b)}} \tag{5}$$

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. 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 claim is proven in our Proposition 5 (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 $f_a(\lambda)$, with $\lambda \in [-\pi, \pi]$, 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 $\lambda \to 0$. $L_f(\cdot)$ is slowly varying at the origin. We 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 (2016) 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.

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*

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

[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).

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.

It should be noted, that 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 for this is 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 (cf. the proofs of Propositions 2 and 4 for details).

As discussed in Leschinski (2016), 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

$$z_t = (y_t - \widehat{y}_{1t})^2 - (y_t - \widehat{y}_{2t})^2 = \widehat{y}_{1t}^2 - \widehat{y}_{2t}^2 - 2y_t(\widehat{y}_{1t} - \widehat{y}_{2t}). \tag{6}$$

Note that even though the forecast objective y_t as well as the forecasts $\widehat{y}_{it}$ in (6), have time index t, the representation is quite versatile. It allows for forecasts 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 $w \times 1$ parameter vector and x_{t-s} is a vector of w 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 (6), 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 (6) 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 (6) 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 long-run 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 the means 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

$$z_t = \widehat{y}_{1t}^{*2} - \widehat{y}_{2t}^{*2} - 2\Big[y_t^*(\mu_1 - \mu_2) + \widehat{y}_{1t}^*(\mu_y - \mu_1) + \widehat{y}_{2t}^*(\mu_y - \mu_2)\Big] - 2\Big[y_t^*(\widehat{y}_{1t}^* - \widehat{y}_{2t}^*)\Big] + const. \quad (7)$$

From (7) 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 (7). Proposition 2 then follows from a case-by-case analysis.

Proposition 2 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 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 (7). Note that the 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 - that is $\mu_1 = \mu_2$. 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 $\xi_y, \xi_i \neq 0$. η_t and ε_{it} are mean zero causal Gaussian long memory processes with parameters d_η and $d_{\varepsilon_{it}}$ 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 analysis. If common long memory according to Assumption 3 is allowed for, this leads to 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 two important special cases: (i) at least one forecast is biased and (ii) all forecasts are unbiased (and $\xi_a = \xi_b$ if a_t and b_t are in a common long memory relationship).

Situation (i) is similar to the first four cases considered in Proposition 2. 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 (7). Since analogous terms to those in the first bracket of (7) 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 therefore obtain the following result.

Proposition 3 (Memory Transmission with Biased Forecasts and CLM). *Under Assumptions 1 and 3, the forecast error loss differential in (6) is* $z_t \sim LM(d_z)$*, where*

$$d_z \geq \begin{cases} d_1, & \text{if } \mu_1 \neq \mu_y \\ d_2, & \text{if } \mu_2 \neq \mu_y \\ d_y, & \text{if } \mu_1 \neq \mu_2. \end{cases}$$

Proof: *See the Appendix.*

Proposition 3 states that the transmission of memory remains the same as in the absence of common long memory, given that the forecasts are biased. As in (7) before, if the forecasts are biased (or have different biases) the memory of the de-meaned series y_t^*, $\widehat{y}_{1t}^*$ and $\widehat{y}_{2t}^*$ dominate that of the other terms. However, if two of those terms appear, it is unclear which one of them is larger - therefore the inequalities in Proposition 3.

The second special case (ii) refers to a situation of unbiasedness similar to the last case in Proposition 2. In addition to that, it is assumed that $\xi_a = \xi_b$, 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 of the common long memory factor x_t driving both series, 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 $y_t - \widehat{y}_{it} = \beta_y - \beta_i + x_t(\xi_y - \xi_i) + \eta_t - \varepsilon_{it}$, so that x_t does not disappear from the linear combination if the scaling parameters ξ_y and ξ_i are different from each other. Hence, we have the following result.

Proposition 4 (Memory Transmission with Unbiased Forecasts and 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 4 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.

The results in Propositions 2, 3 and 4 show that long memory can be transmitted from forecasts or the forecast objective to the forecast error loss differentials. This situation can arise naturally in many practical situations. First, of course the forecast objective might be a long memory time series. Second, from Proposition 3 in Chambers (1998), forecasts that are based on linear combinations - such as predictive regressions - exhibit long memory if they include a long memory variable.

Our results also show that the biasedness of the forecasts plays an important role for the transmission of dependence to the loss differentials. In practical situations, it might be overly restrictive to impose exact unbiasedness (under which memory would be reduced according to Proposition 4). 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).

It is well established that estimation errors might imply biased forecasts. 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.

Finally, even optimal forecasts can be strongly persistent for long forecast horizons. It is well known that the forecast errors of an optimal h-step-ahead forecast follow an $\mathrm{MA}(h-1)$ process. The coefficients of this process are given by the first $h-1$ coefficients in the $\mathrm{MA}(\infty)$ representation of the objective series y_t. Therefore, forecast errors of the h-step forecast for $h \to \infty$ have long memory if the underlying series y_t has long memory as well. For larger forecast horizons long memory processes are therefore a better approximation to the true dependence structure than short memory processes.

3.4 Asymptotic and Finite-Sample Behaviour under Long Memory

After confirming that forecast error loss differentials can exhibit long memory, 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 5 (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 $T \to \infty$.*

Proof: *See the Appendix.*

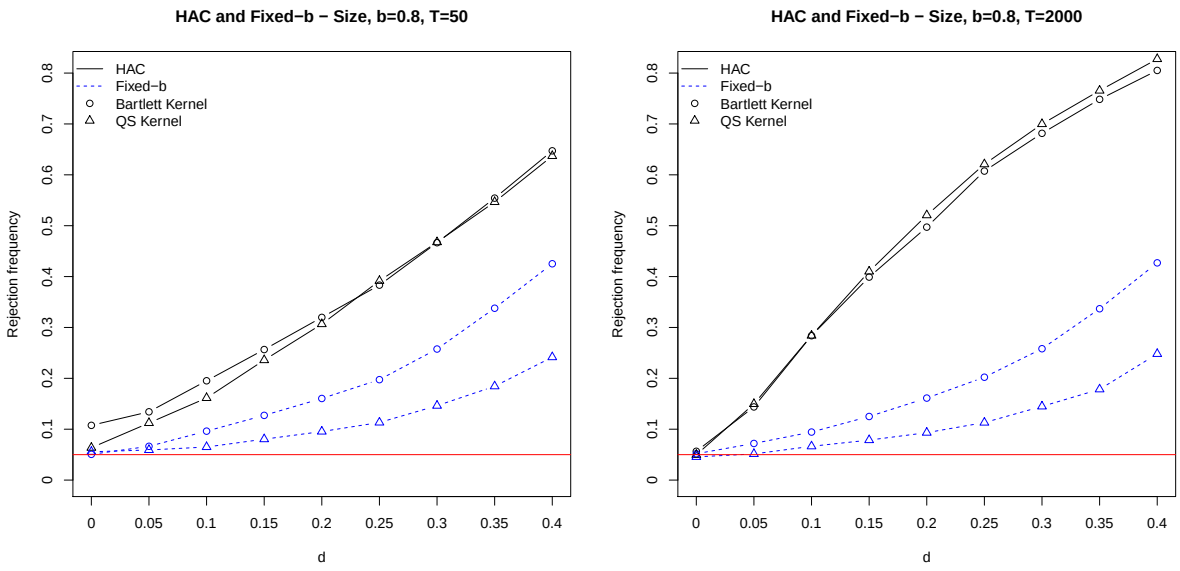

Figure 1: Size of the t_{HAC}- and t_{FB}-tests with $T \in \{50, 2000\}$ for different values of the memory parameter d.

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, both tests are over-sized. These findings motivate the use of long memory robust procedures. Continuing with the results for $T = 2000$ in the right panel of Figure 1, we observe similar findings in general. We note that for the short memory case, size distortions arising from a too small sample size vanish.

[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.

All tests statistics are well behaved for $d = 0$. On the contrary, size distortions are stronger with an increasing sample size, although the magnitude of additional distortion is moderate. This feature can be attributed to the slow divergence rate (as given in the proof of Proposition) 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, it is necessary to 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). Note that 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 $T \to \infty$

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

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 process, estimation of the long-run variance additionally hinges on the estimation of d. The

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

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 $m_d = [T^q]$, where $0 < q < 1$ denotes a generic bandwidth parameter. This estimator is given by

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

where $U_T(d) = \log\left(\frac{1}{m_d}\sum_{j=1}^{m_d} j^{2d} I_T(\lambda_j)\right) - \frac{2d}{m_d}\sum_{j=1}^{m_d}\log j$. Many other estimation approaches (e.g. log-periodogram estimation, etc.) would be a possibility as well.

Next, b_0 can be estimated consistently by

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

where $I_T(\lambda_j)$ 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 $\lambda_j = 2\pi j/T$ are the Fourier frequencies for $j = 1, ..., \lfloor T/2 \rfloor$. Here, $\lfloor \cdot \rfloor$ denotes the largest integer smaller than its argument. 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 $m = [T^{4/5}]$ 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]

[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.

5 Monte Carlo Study

In this section we analyze the finite-sample performance of the procedures discussed above by means of a simulation study. As in our motivating example, we conduct all size and power simulations for the t_{MAC}- and t_{EFB}-tests with $M = 5000$ Monte Carlo repetitions and the nominal significance level is set to 5%. For both tests, the plug-in estimation of d is done via local Whittle (LW) with $m_d = \lfloor T^{0.65} \rfloor$ which is similar to the simulation setup in Abadir et al. (2009). In the case of the extended fixed-b approach, we consider the Bartlett and the Modified Quadratic Spectral (MQS) kernel as used in Politis and McElroy (2009) and McElroy and Politis (2012).[5]

Note that even though the theoretical results in Section 3 are based on assumptions on the forecasts and the forecast objective, the modified DM tests proposed in Section 4 are based solely on assumptions on the time series properties of the loss differentials. Since these tests are the subject of this Monte Carlo study, we also take this perspective for the simulation design and generate the loss differential series z_t directly from standard time series models.

The results reported below are generated for the following two DGPs. DGP1 is a fractional Gaussian white noise process with memory parameter $d = \{0, 0.05, 0.1, ..., 0.4\}$, while DGP2 contains an additional first-order autoregressive component with parameter $\phi = 0.6$.

If the loss differential series has zero mean, this represents a situation where both forecasts are equally good. For non-zero means one of the forecasts outperforms the other. Since the DM test is essentially a test on the mean, the results presented below can not only be interpreted with regard to forecast comparisons. Instead, they can also be considered as a general comparison of size and power between statistics using the MAC estimator and tests employing the extended fixed-b asymptotics. To the best of our knowledge, such a comparison has not been conducted in the existing literature before.

In regard of the fact that optimal forecasts are MA processes the attentive reader might wonder why the results presented do not include MA dynamics. However, the derivative of the spectral density of MA processes in the vicinity of the zero frequency tends to be much smaller than that of AR processes, so that the spectral density at the origin is more flat and has a less severe effect on the finite-sample performance of the estimators for the long memory parameters. We therefore decide to present the results for the situation that is more challenging for the methods employed, but additional results under first-order MA dynamics (with MA parameter $\theta = 0.6$) are available in our online appendix. In addition to that, the important special case of optimal one-step-ahead forecasts is represented by DGP1 for $d = 0$.

Figure 2 shows a comparison of the size of both tests for different degrees of long memory and sample sizes of $T \in \{250, 2000\}$.[6] In the case of DGP1 (black solid lines), tests are liberal and the size tends to increase with increasing d. The t_{EFB}-statistic obtained with the MQS kernel gives the best size control, whereas the t_{MAC}-statistic shows the highest rejection frequencies. In

[5] The MQS kernel is a modified version of the usual QS kernel used in Kiefer and Vogelsang (2005), but restricted to $x \in [-1, 1]$. The kernel is given by $k(x) = 3(\sin(\pi x)/(\pi x) - \cos(\pi x))/(\pi x)^2$ for $x \in [-1, 1]$ and $k(x) = 0$ for $|x| > 1$, where $x = j/B$, if the kernel is employed for the long-run variance estimation as in (3). Further kernels, including flat-top tapers, are analyzed as well but yield slightly inferior results to those reported here.

[6] Additional simulation results for $T \in \{50, 1000\}$ are reported in our online appendix.

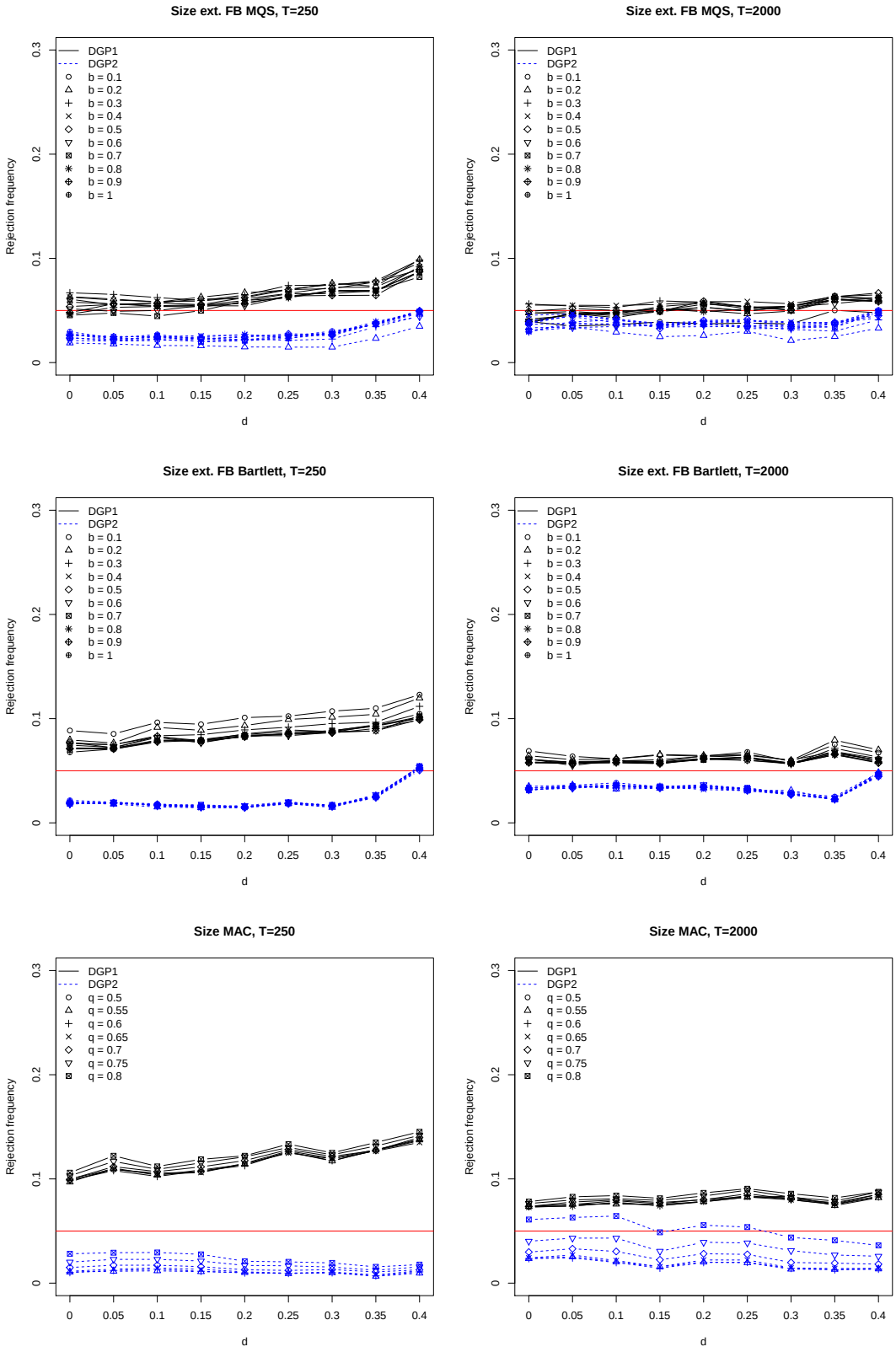

Figure 2: Size of the t_{MAC}- and t_{EFB}-statistics for different degrees of long memory d, sample sizes $T \in \{250, 2000\}$ and bandwidth parameters q and b.

larger samples of $T = 2000$ observations the dependence of the size on d is reduced and both tests approach their nominal significance level. However, for both sample sizes the t_{EFB}-statistics are

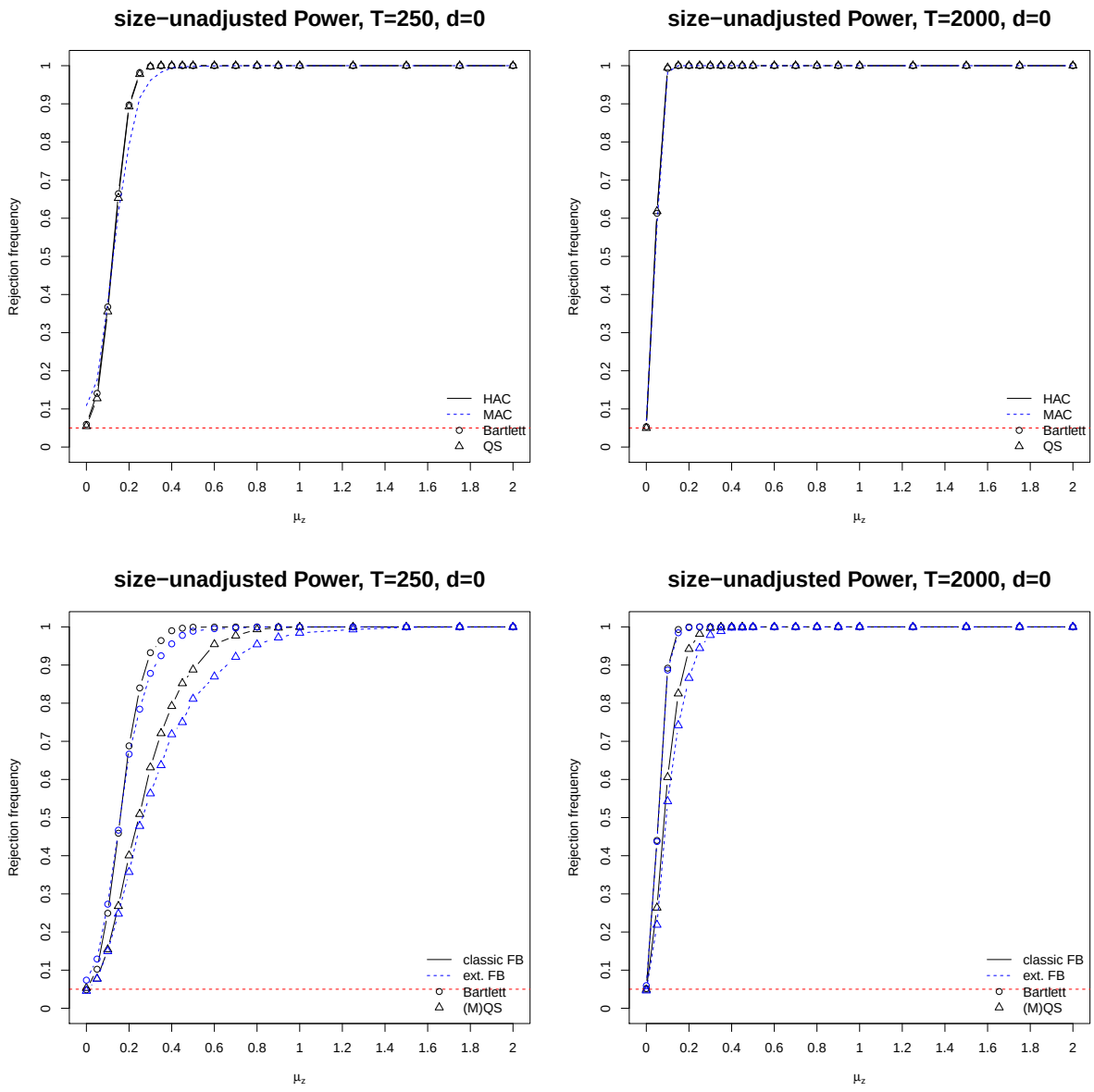

Figure 3: Power comparison of the robust statistics t_{EFB} and t_{MAC} with their short memory counterparts when $d = 0$.

notably closer to their nominal level of 5% than the MAC-based statistic and among the t_{EFB}-statistics, the one obtained with the MQS kernel performs best. As observed by Kiefer and Vogelsang (2005), there is a trade-off in terms of size and power in the choice of the bandwidth parameter b. Larger bandwidths generally improve the size and reduce the power. However, as can be seen from the results below, the kernel choice has a more severe effect than the bandwidth choice. Especially in larger samples the size is nearly identical for all bandwidths.

DGP2 (blue dashed lines) contains short memory influences and the results shown here are obtained with an autoregressive coefficient of $\phi = 0.6$. Interestingly, in the presence of short memory components, the results change notably. Already in samples of $T = 250$ observations both tests are conservative. For large values of d, we can observe that the downward size distortions of the t_{EFB}-test vanish. In large samples, all tests show better empirical size properties,

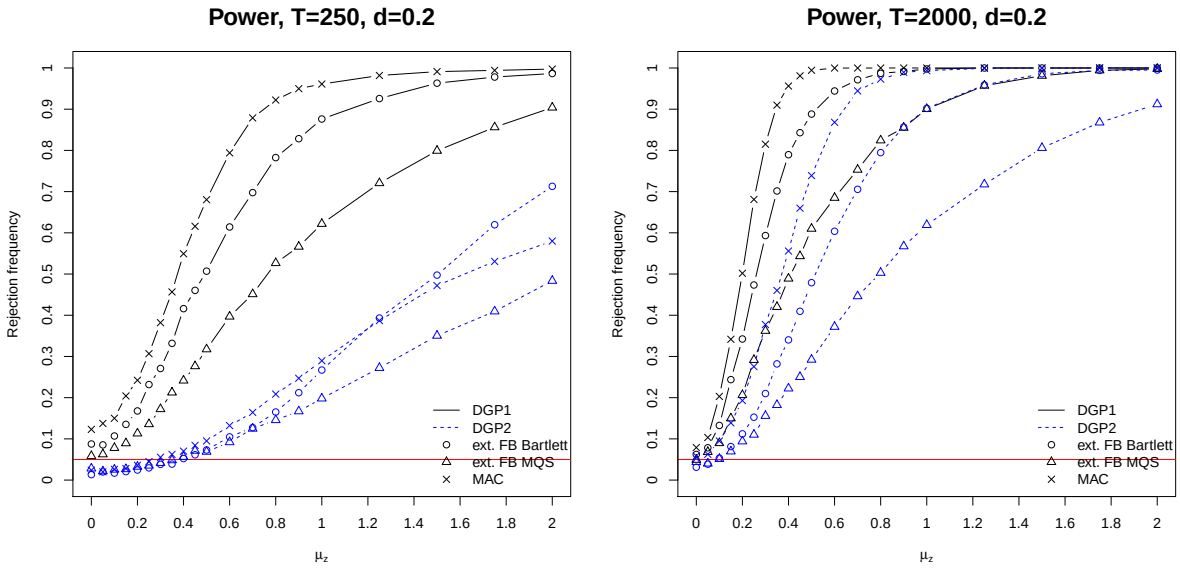

Figure 4: Power comparison of the t_{MAC}- and t_{EFB}-statistics for $d = 0.2$ and sample sizes $T \in \{250, 2000\}$.

as expected. Further simulations considering moving average components are conducted and results are reported in our online appendix. Qualitatively, this does not alter the findings, even though the conservativeness of the procedure becomes stronger with increasing ϕ and moving average components tend to have a less severe impact compared to autoregressive components for the reasons discussed above.

Our simulation study suggests that a bandwidth choice of $b = 0.8$ provides a good balance in the size-power trade-off under both DGPs, for both kernels, and for all considered memory parameters. Concerning the MAC estimator, the MSE-optimal choice $m_{opt} = \lfloor T^{0.8} \rfloor$ derived in Abadir et al. (2009) indeed provides the best results under DGP2. In this situation, it gives a size close to the nominal level and is better than that of the t_{EFB}-test. However, in the case of DGP1 the bandwidth $m = \lfloor T^q \rfloor$ with $q = 0.7$ seems more adequate which can also observed in the simulation study of Abadir et al. (2009).

In a next step, we consider the potential losses in power arising from the use of the robust t_{EFB}- and t_{MAC}-statistics when the additional flexibility is not needed, because the series is short memory ($d = 0$). With regard to the previous results, we choose the bandwidth parameter of $b = 0.8$ for the extended fixed-b and $m = \lfloor T^{0.7} \rfloor$ ($m = \lfloor T^{0.8} \rfloor$) for the MAC approach under DGP1 (DGP2). Results are presented in Figure 3. Since it is our objective to evaluate the potential loss in power if one would generally use memory robust tests in practice, we consider size-unadjusted power here. We compare the t_{HAC}- and t_{MAC}-tests in the top row and the t_{FB}- with the t_{EFB}-statistics in the bottom row of Figure 3 with $T \in \{250, 2000\}$ and DGP1, setting $d = 0$. Although some power loss can be observed as expected, the cost of using the long memory robust procedures is more than acceptable, already for $T = 250$.

Finally, we analyze the power of the t_{MAC}- and t_{EFB}-statistics under both DGPs, for the case of $d = 0.2$. This setup matches the memory orders in our empirical application (cf. Section 6) closely. We choose the same bandwidth parameters as before to enhance comparability. To

control for the increase in the variance of the process (which depends on the memory parameter d), each loss differential series is standardized before the mean (μ_z) is added and the respective test is applied. The results are shown in Figure 4. As expected, the power increases with the sample size.

With regard to the ranking, we observe that in case of DGP1 the t_{MAC}-statistic clearly outperforms the t_{EFB}-statistics among which the one obtained using a Bartlett kernel performs best. The t_{EFB}-statistic obtained with the MQS kernel, on the other hand, has the lowest power under both DGPs. Since the t_{MAC}-statistic is clearly more liberal than its two competitors, we also provide size-adjusted power curves in Figure 9 (and 16, for MA dynamics) in our online appendix. Due to the different employment of the d estimator in both methods, such a comparison is only valid for known d. For both DGPs, it can be clearly seen that the power advantages of the t_{MAC}-statistic go beyond the effect of the upward size distortion.

By comparing the results for DGP1 with those of DGP2 in Figure 4, one can observe that the power of both tests suffers if short memory components are present. Different from the size effect of the short memory dynamics discussed above, simulations with known d show that this cannot be explained by the effect of autoregressive dynamics on the estimation of d alone. Instead, the presence of short memory dynamics increases the finite-sample variance of the estimated means - similar to the effect of an increase in d.

For robustification of the procedures against the effect of short memory dynamics discussed above, one could consider to apply the adaptive local polynomial Whittle (ALPW) estimator of Andrews and Sun (2004). Figures 5 and 6 (and 13 and 14, for MA dynamics) in our online appendix shows the results of this exercise. In small samples, the size obtained using the ALPW estimator becomes similarly liberal for all procedures and both DGPs. In larger samples of $T = 250$ and beyond, all tests reach a satisfactory size, however, the size of the t_{EFB}-statistic using the MQS kernel remains the best and the t_{MAC}-statistic performs better if a smaller bandwidth, say $m = \lfloor T^{0.55} \rfloor$ is used. The power, on the other hand, is remarkably reduced and the t_{EFB}-statistic using the Bartlett kernel has the highest power for sample sizes of $T \in \{50, 250\}$. For larger samples, however, the former ranking is reestablished suggesting a small sample effect.

We find that the t_{EFB}-tests generally provide better size control than the t_{MAC}-test, whereas the latter has better power properties. Among the extended fixed-b procedures, the MQS kernel has better size but less power compared to the Bartlett kernel. In presence of short memory dynamics both procedures become quite conservative. This effect can be mitigated if the ALPW estimator is employed for the plug-in estimation of the memory parameter d. However, this comes at the cost of an additional loss in power which might be attributed to the increased variance of the estimator, partly also due to the automatic bandwidth selection approach.

Since there is no dominant procedure in terms of size control and power, we conclude that it is beneficial for forecast comparisons in practice to consider both statistics and to compare the outcomes. In our empirical applications to realized volatility in the next section we consider such comparisons. In general, our conclusions also apply to other inference problems involving the sample mean.

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., e.g., 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.

As empirical applications we therefore 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, than following Andersen et al. (2001) and Barndorff-Nielsen and Shephard (2002) the daily realized variance is defined as

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

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

$$\ln RV_t^{(h)} = \alpha + \rho_{22} \ln RV_{t-h}^{(22)} + \rho_5 \ln RV_{t-h}^{(5)} + \rho_1 \ln RV_{t-h}^{(1)} + \varepsilon_t \ , \tag{8}$$

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 that highlighted 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 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, if these are realized. Therefore, the quadratic variation of the cumulative return process can be 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.[7] 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 component C_t and the jump component J_t as

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

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 that find 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. Here,

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

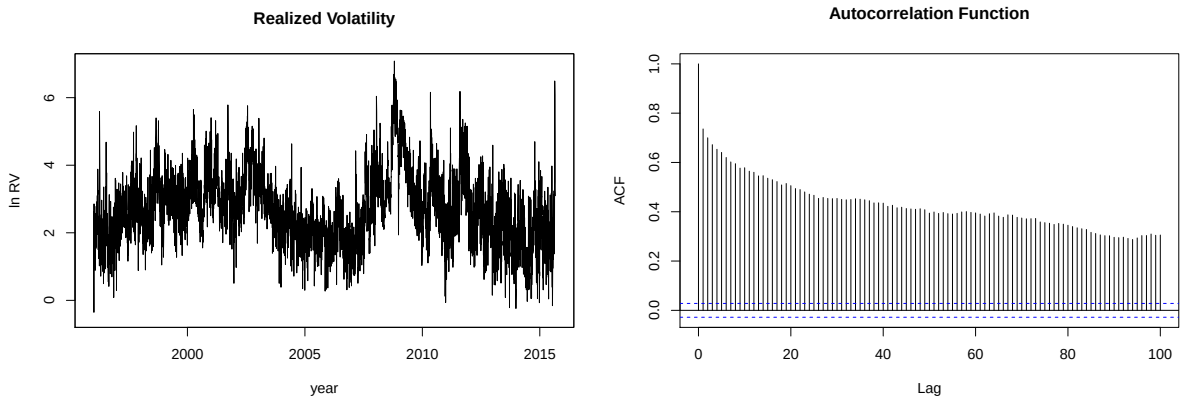

Figure 5: Daily log-realized volatility of the S&P500 index and their autocorrelation function.

we use the HAR-RV-TCJ model that is studied in Bekaert and Hoerova (2014):

$$\begin{aligned}\ln RV_t^{(h)} &= \alpha + \rho_{22} \ln C_{t-h}^{(22)} + \rho_5 \ln C_{t-h}^{(5)} + \rho_1 \ln C_{t-h}^{(1)} \\ &\quad + \varpi_{22} \ln\left(1 + J_{t-h}^{(22)}\right) + \varpi_5 \ln\left(1 + J_{t-h}^{(5)}\right) + \varpi_1 \ln\left(1 + J_{t-h}^{(1)}\right) + \varepsilon_t \,. \end{aligned} \tag{9}$$

The daily log-realized variance series ($\ln RV_t$) is depicted in Figure 5.[8] 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 5, 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 1. Local Whittle estimates ($\widehat{d}_{LW}$) exceed 0.5 slightly and thus indicate 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 to low-frequency contaminations. As one can see, the memory estimates are fairly stable and the Qu 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. We therefore also report results of the local polynomial Whittle plus noise (LPWN) estimator of Frederiksen et al. (2012). Similar to the ALPW estimator of Andrews and Sun (2004), the LPWN estimator reduces the bias due to short memory dynamics by approximating the log-spectral density of the short memory component with a polynomial, but it additionally includes a second polynomial to account for the downward bias induced by perturbations. As one can see, 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.

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

q	$\widehat{d}_{LW}$	$\widehat{d}_{HP}$	s.e.	W	$\widehat{d}_{(0,0)}$		$\widehat{d}_{(1,0)}$		$\widehat{d}_{(1,1)}$	
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 1: 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 (W statistic) 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. For details, see Frederiksen et al. (2012).

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, 3 and 4 can occur.

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 6. As one can see, 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 $VP_t = \ln(VIX_t^2/12) - \ln RV_{t+22}^{(22)}$ and displayed on the right hand side of Figure 6. The graph clearly suggests that the VIX tends to overestimate the realized variance and the sample average of the variance risk premium is 0.623. Furthermore, the linear combination of realized and 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 VP_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

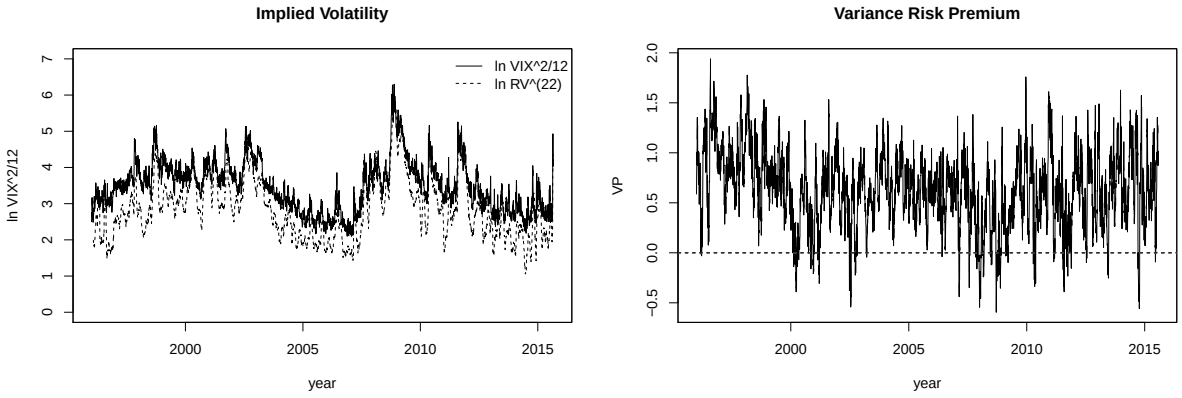

Figure 6: Log squared implied volatility and log cumulative realized volatility of the S&P 500 (left panel) and variance risk premium $VP_t = \ln(VIX_t^2/12) - \ln RV_{t+22}^{(22)}$ (right panel).

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 linearly decaying coefficients. We therefore base all our robust statistics on the local polynomial Whittle plus noise (LPWN) estimator of Frederiksen et al. (2012) discussed above.[9] Since Chen and Ghysels (2011) and Corsi and Renò (2012) show that leverage effects improve 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 2 reports the results. Models are estimated using a rolling window of $T_w = 1000$ observations.[10] This implies that the forecast window contains 3883 observations.[11] All DM tests are conducted with one-sided alternatives. We test that a more complex model outperforms its parsimonious version. For the sake of a better comparability, all kernel-based tests use the Bartlett kernel. In accordance with the previous literature, the t_{DM}-statistic is implemented using an MA approximation with 44 lags for the forecast horizon of 22 days, c.f. for instance Bekaert and Hoerova (2014). For the t_{HAC}-statistic we use an automatic bandwidth selection procedure and

[9] We choose $R_y = 1$ and $R_w = 0$ concerning the polynomial degrees and a bandwidth $m_d = \lfloor T^{0.8} \rfloor$ (see Frederiksen et al. (2012) for details on the estimator).

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

[11] Additional simulation results for a sample size of 4000 observations are available in our online appendix. The results are generally in line with those for 2000 observations.

Models	Summary statistics					Short memory inference			Long memory inference						
									t_{MAC}			t_{EFB}			
Model vs. Model+VIX	$\overline{z}/\widehat{\sigma}_z$	$MSE1$	$MSE2$	$\widehat{d}_{LW}$	$\widehat{d}_{LPWN}$	t_{DM}	t_{HAC}	t_{FB}	0.7	0.75	0.8	0.2	0.4	0.6	0.8
HAR-RV	0.135	0.292	0.269	**0.219***	**0.234***	**2.968**	**3.032**	**2.494**	0.929	1.038	1.188	2.494 (3.404)	2.754 (4.064)	2.985 (4.750)	2.849 (5.388)
HAR-RV-TCJ	0.109	0.285	0.268	**0.175***	0.138	**2.421**	**2.455**	**2.097**	1.397	1.610	**1.892**	2.097 (2.610)	2.503 (3.154)	2.889 (3.693)	2.724 (4.228)
HAR-RV-TCJ-L	0.082	0.282	0.269	**0.182***	0.163	**1.784** (1.645)	**1.786** (1.645)	1.819 (2.092)	0.889	1.016 (1.645)	1.192	1.819 (3.404)	2.153 (4.064)	2.430 (4.750)	2.317 (5.388)

Table 2: 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 MSEs of the models 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.

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 2 reveals that the forecast error loss differentials have long memory with d parameters between 0.138 and 0.234. The results are very similar for the local Whittle and the LPWN estimator. Standard DM statistics (t_{DM}, t_{HAC} and t_{FB}) reject the null hypothesis of equal predictive ability, thereby confirming the findings in the previous literature. However, if the memory robust statistics in the right panel of Table 2 are taken into account, all evidence for a superior predictive ability of models including the VIX vanishes. Therefore, the previous rejections might be spurious and reflect the theoretical findings in Proposition 5. 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 comparison, we also consider the QLIKE loss function

$$g(y_t, \widehat{y}_t) = \log(\widehat{y}_t) + \frac{y_t}{\widehat{y}_t}$$

in addition to the MSE. The motivation for this 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. We therefore also consider QLIKE, even though our theoretical results in Section 3 do not apply to this loss function.

Results are reported in Table 3. They suggest that the average standardized forecast error loss differentials are positive and similar in magnitude to the MSE comparison in Table 2. 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,

Models		Summary statistics				Short memory inference			Long memory inference						
										t_{MAC}			t_{EFB}		
Model vs. Model+VIX	$\overline{z}/\widehat{\sigma}_z$	$QLIKE1$	$QLIKE2$	$\widehat{d}_{LW}$	$\widehat{d}_{LPWN}$	t_{DM}	t_{HAC}	t_{FB}	0.7	0.75	0.8	0.2	0.4	0.6	0.8
HAR-RV	0.152	2.025	2.023	**0.234***	**0.203**	**3.071**	**3.251**	**2.758**	1.299	1.453	**1.657**	2.758	3.024	3.161	3.068
												(3.404)	(4.064)	(4.750)	(5.388)
HAR-RV-TCJ	0.129	2.025	2.023	**0.188***	0.133	**2.720**	**2.806**	**2.833**	**1.720**	**1.971**	**2.288**	**2.833**	**3.236**	3.446	3.367
												(2.610)	(3.154)	(3.693)	(4.228)
HAR-RV-TCJ-L	0.101	2.024	2.023	**0.186***	0.103	**2.065**	**2.053**	**2.517**	1.574	**1.837**	**2.167**	2.517	2.869	3.073	3.029
						(1.645)	(1.645)	(2.092)		(1.645)		(2.610)	(3.154)	(3.693)	(4.228)

Table 3: 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 2.

the null hypothesis of pairwise equal predictive ability amongst the models is rejected in all cases. This is also in line with the previous results.

Turning to long memory-robust statistics, we find somewhat different results, especially for the symmetric HAR-RV-TCJ model including jumps. Here, we mostly observe rejections of equal predictive ability in favor of the inclusion of the VIX. This is likely to be due to the asymmetry of the QLIKE loss function. For other versions of the HAR model, the evidence against the null hypothesis is weaker and thus mainly in line with our previous results.

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 (9) leads to a significant improvement in forecast performance compared to the standard HAR-RV-model (8) from a purely out-of-sample perspective.

The continuous components and jump components - separated using the approach described above - are shown in Figure 7. 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 log-continuous component closely follows that of the log-realized volatility series.

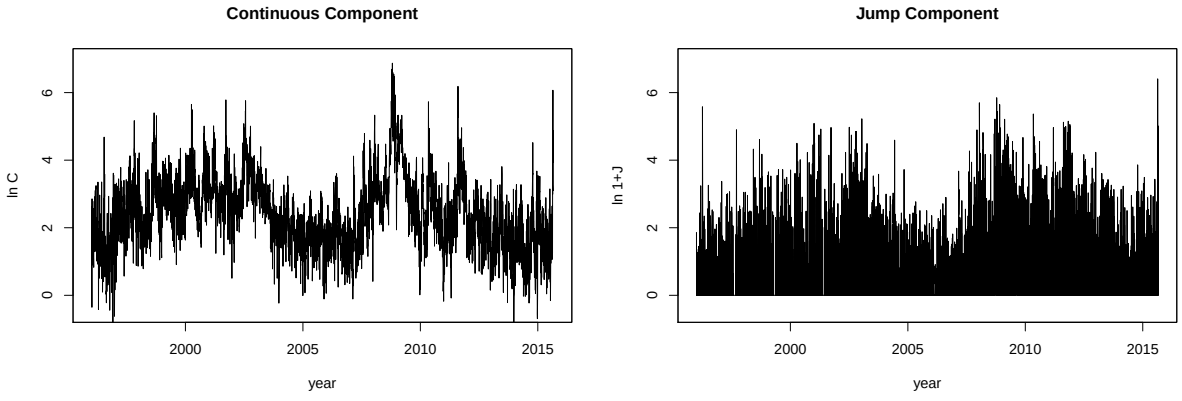

Figure 7: Log continuous component $\ln C_t$ and jump component $\ln(1 + J_t)$ of RV_t.

Models	Summary statistics					Short memory inference			Long memory inference						
									t_{MAC}			t_{EFB}			
HAR-RV vs.	$\overline{z}/\widehat{\sigma}_z$	MSE_1	MSE_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-TCJ, $h=1$	0.122	0.409	0.375	**0.094***	0.127	**6.932**	**7.631**	**3.995**	**3.243**	**3.144**	**3.091**	**3.995** (2.610)	**4.068** (3.154)	**4.468** (3.693)	**4.947** (4.228)
HAR-RV-TCJ, $h=5$	0.092	0.263	0.247	**0.072**	0.009	**3.666**	**3.790**	**2.789**	**3.620**	**3.853**	**4.277**	**2.789** (2.050)	**3.981** (2.522)	**5.093** (2.975)	**5.848** (3.386)
HAR-RV-TCJ, $h=22$	0.045	0.292	0.285	**0.359***	**0.343***	0.776 (1.645)	0.912 (1.645)	0.666 (2.092)	0.140	0.152 (1.645)	0.171	0.666 (4.701)	0.925 (5.551)	1.064 (6.413)	1.164 (7.281)

Table 4: Separation of Continuous and Jump Components (evaluated under MSE loss). 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 MSEs of the models 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.

Table 4 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 this literature. All other specifications are the same as before. As one can see, the standard tests (t_{DM}, t_{HAC} and t_{FB}) agree upon rejection of the null hypothesis of equal predictive ability 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 $\widehat{d}_{LW} = 0.094$ and $\widehat{d}_{LW} = 0.070$ which is qualitatively similar to the results obtained using the LPWN estimator. Therefore, the rejections of equal predictive accuracy obtained using standard tests might be spurious due to the neglected effect of long range dependence. Nevertheless, the improvement in forecast accuracy is large enough, so that the long memory robust t_{MAC}- and t_{EFB}-statistics reject across the board for $h = 1$ and $h = 5$.

Models	Summary statistics					Short memory inference			Long memory inference						
									t_{MAC}			t_{EFB}			
HAR-RV vs.	$\overline{z}/\widehat{\sigma}_z$	$QLIKE_1$	$QLIKE_2$	$\widehat{d}_{LW}$	$\widehat{d}_{LPWN}$	t_{DM}	t_{HAC}	t_{FB}	0.7	0.75	0.8	0.2	0.4	0.6	0.8
HAR-RV-TCJ, $h=1$	0.066	1.932	1.930	0.044	0.009	**4.409**	**4.080**	**2.834**	**4.193**	**3.944**	**3.823**	**2.834** (2.050)	**3.012** (2.522)	**3.368** (2.975)	**3.731** (3.386)
HAR-RV-TCJ, $h=5$	0.035	1.986	1.986	**0.089***	0.016	1.384	1.422	1.288	1.326	1.387	1.517	1.288 (2.050)	1.820 (2.522)	2.577 (2.975)	2.795 (3.386)
HAR-RV-TCJ, $h=22$	−0.007	2.025	2.025	**0.425***	**0.382***	−0.121 (1.645)	−0.141 (1.645)	−0.113 (2.092)	−0.016	−0.017 (1.645)	−0.019	−0.113 (7.486)	−0.147 (8.692)	−0.161 (9.974)	−0.169 (11.417)

Table 5: Separation of Continuous and Jump Components (evaluated under QLIKE loss). See the notes for Table 4.

When considering the QLIKE loss function as an alternative to the MSE in Table 5, we find evidence against the null hypothesis for the case of short-term forecasting, but no for the weekly and monthly horizon. We can therefore confirm that the separation of continuous and jump components indeed improves the forecast performance on daily horizons.

7 Conclusion

This paper deals with forecast evaluation under long range dependence. We show in Section 3 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. 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 has long memory.

Two methods for robustification of DM tests against long memory are discussed in Section 4 - the MAC estimator of Robinson (2005) and Abadir et al. (2009), as well as the extended fixed-b approach of McElroy and Politis (2012).

The finite sample performance of both of these methods is studied using Monte Carlo simulations. While the extended fixed-b approach allows a better size control, the MAC performs better in terms of power. With regard to kernel and bandwidth choices for the t_{EFB}-statistic, we find that $b = 0.8$ gives good results and that the kernel choice has a larger impact on the size and power of the procedure than the bandwidth selection in absence of short-run dynamics. In general, the MQS kernel gives a better size control, whereas the Bartlett kernel is superior in terms of power. An important issue remains the impact of short memory dynamics on the plug-in estimation of the memory parameter. However, our results using the ALPW estimator of Andrews and Sun (2004) indicate that bias-corrected local Whittle estimators successfully improve the results - at least in larger samples. As to be expected, this comes at the price of a power loss.

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. We therefore consider this series in our empirical application. 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.

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.

Andrews, D. W. and Sun, Y. (2004). Adaptive local polynomial whittle estimation of long-range dependence. *Econometrica*, 72(2):569–614.

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.

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 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.

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.

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. (2016). On the memory of products of long range dependent time series. Technical Report 569, Leibniz University of Hannover.

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

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.

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

Politis, D. and McElroy, T. S. (2009). Fixed-b asymptotics for the studentized mean from time series with short, long or negative memory. *Department of Economics, UCSD*.

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.

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 (6) can be re-expressed as*

$$\begin{aligned}
z_t =& -2y_t(\widehat{y}_{1t} - \widehat{y}_{2t}) + \widehat{y}_{1t}^2 - \widehat{y}_{2t}^2 \\
=& -2(y_t^* + \mu_y)[\widehat{y}_{1t}^* + \mu_1 - \widehat{y}_{2t}^* - \mu_2] + (\widehat{y}_{1t}^* + \mu_1)^2 - (\widehat{y}_{2t}^* + \mu_2)^2 \\
=& -2\{y_t^*\widehat{y}_{1t}^* + \mu_1 y_t^* - y_t^*\widehat{y}_{2t}^* - y_t^*\mu_2 + \mu_y\widehat{y}_{1t}^* + \mu_y\mu_1 - \widehat{y}_{2t}^*\mu_y - \mu_2\mu_y\} \\
& + \widehat{y}_{1t}^{*2} + 2\widehat{y}_{1t}^*\mu_1 + \mu_1^2 - \widehat{y}_{2t}^{*2} - 2\widehat{y}_{2t}^*\mu_2 - \mu_2^2 \\
=& \underbrace{-2[y_t^*(\mu_1 - \mu_2) + \widehat{y}_{1t}^*(\mu_y - \mu_1) - \widehat{y}_{2t}^*(\mu_y - \mu_2)]}_{I} - \underbrace{2[y_t^*(\widehat{y}_{1t}^* - \widehat{y}_{2t}^*)]}_{II} + \underbrace{\widehat{y}_{1t}^{*2} - \widehat{y}_{2t}^{*2}}_{III} + const. \quad (10)
\end{aligned}$$

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 (2016), 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 (10). 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 (2016). We thus have for $i \in \{1, 2\}$

$$y_t^*\widehat{y}_{it}^* \sim \begin{cases} LM(\max\{d_y + d_i - 1/2, 0\}), & \text{if } S_{y,\widehat{y}_i} \neq 0 \\ LM(d_y + d_i - 1/2), & \text{if } S_{y,\widehat{y}_i} = 0 \end{cases} \quad (11)$$

$$\text{and} \qquad \widehat{y}_{it}^{*2} \sim LM(\max\{2d_i - 1/2, 0\}). \quad (12)$$

Further note that

$$d_y > d_y + d_i - 1/2 \;\; \text{and} \;\; d_i > d_y + d_i - 1/2 \quad (13)$$

and

$$d_i > 2d_i - 1/2, \quad (14)$$

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

Using these properties, we can determine the memory d_z in (10) 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 (13) and (14) 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 (10), but the two forecasts $\widehat{y}_{1t}$ and $\widehat{y}_{2t}$ remain. From (13) and (14), 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 (13) and (14), 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 (10). 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 (11) and (12). Furthermore, the memory of the squared series given in (12) is always non-negative from Corollary 1 in Leschinski (2016) 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*

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

□

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

$$\mu_i = \beta_i + \xi_i \mu_x \tag{15}$$

and

$$\mu_y = \beta_y + \xi_y \mu_x. \tag{16}$$

We can now re-express the forecast error loss differential z_t *in (10) 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*

$$\begin{aligned} z_t = &-2\{y^*_t(\mu_1 - \mu_2) + x^*_t[\xi_1(\mu_y - \mu_1) - \xi_2(\mu_y - \mu_2)] + x^*_t y^*_t(\xi_1 - \xi_2) - x^*_t(\xi_1\varepsilon_{1t} - \xi_2\varepsilon_{2t}) \\ &+ \varepsilon_{1t}(\mu_y - \mu_1) - \varepsilon_{2t}(\mu_y - \mu_2) + \mu_x(\varepsilon_{1t}\xi_1 - \varepsilon_{2t}\xi_2) + y^*_t(\varepsilon_{1t} - \varepsilon_{2t})\} \\ &+ x^{*^2}_t(\xi_1^2 - \xi_2^2) + \varepsilon_{1t}^2 - \varepsilon_{2t}^2 + 2\mu_x(\varepsilon_{1t}\xi_1 - \varepsilon_{2t}\xi_2) + const. \end{aligned} \tag{17}$$

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

$$\begin{aligned} z_t = &-2\{x^*_t[(\mu_y - \mu_1)\xi_1 + \xi_y(\mu_1 - \mu_2)] - \widehat{y}^*_{2t}[\mu_y - \mu_2] - \xi_y x^*_t \widehat{y}^*_{2t} + x^*_t[\varepsilon_{1t}(\xi_y - \xi_1) + \xi_1\eta_t] \\ &+ \varepsilon_{1t}(\xi_y\mu_x - \mu_1) + \eta_t(\mu_1 - \mu_2) + \varepsilon_{1t}\eta_t - \widehat{y}^*_{2t}\eta_t\} \\ &- (2\xi_1\xi_y - \xi_1^2)x^{*^2}_t + \varepsilon_{1t}^2 - \widehat{y}^{*^2}_{2t} - 2\beta_y\varepsilon_{1t} + const. \end{aligned} \tag{18}$$

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:*

$$\begin{aligned} z_t = &-2\big\{x_t^*[\xi_1(\mu_y-\mu_1)-\xi_2(\mu_y-\mu_2)+\xi_y(\mu_1-\mu_2)] \\ &+ x_t^*[(\xi_y-\xi_1)\varepsilon_{1t}-(\xi_y-\xi_2)\varepsilon_{2t}+(\xi_1-\xi_2)\eta_t] \\ &+ x_t^{*2}[\xi_y(\xi_1-\xi_2)-\frac{1}{2}(\xi_1^2-\xi_2^2)] \\ &+ \varepsilon_{1t}(\mu_y-\mu_1)-\varepsilon_{2t}(\mu_y+\mu_2)+\mu_x(\xi_1\varepsilon_{1t}+\xi_2\varepsilon_{2t})+\eta_t(\varepsilon_{1t}-\varepsilon_{2t})+\eta_t[\mu_1-\mu_2]\big\} \\ &+ \varepsilon_{1t}^2-\varepsilon_{2t}^2+2\mu_x(\xi_1\varepsilon_{1t}-\xi_2\varepsilon_{2t})+const. \end{aligned} \tag{19}$$

As in the proof of Proposition 2, we can now determine the memory orders of z_t in (17), (18) and (19) by first considering the memory of each term in each of the linear combinations and then by applying Proposition 3 of Chambers (1998) thereafter. Note, however, that

$$y_t^*(\mu_1-\mu_2)+x_t^*[\xi_1(\mu_y-\mu_1)-\xi_2(\mu_y-\mu_2)] \;\; \textit{in (17)},$$

$$x_t^*[(\mu_y-\mu_1)\xi_1+\xi_y(\mu_1-\mu_2)]-\widehat{y}_{2t}^*(\mu_y-\mu_2) \;\; \textit{in (18)}$$

and

$$x_t^*[\xi_1(\mu_y-\mu_1)-\xi_2(\mu_y-\mu_2)+\xi_y(\mu_1-\mu_2)] \;\; \textit{in (19)}$$

have the same structure as

$$y_t^*(\mu_1-\mu_2)+\widehat{y}_{1t}^*(\mu_y-\mu_1)-\widehat{y}_{2t}^*(\mu_y-\mu_2) \;\; \textit{in (10)}$$

and that all of the other non-constant terms in (17), (18) and (19) are either squares or products of demeaned series, so that their memory is reduced according to Proposition 1 from Leschinski (2016). 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 (13) and (14), 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. □

Proof (Proposition 4). *First note that under the assumptions of Proposition 3, (17) is reduced to*

$$\begin{aligned} z_t &= -2\{-x_t^*(\xi_1\varepsilon_{1t} - \xi_2\varepsilon_{2t}) + y_t^*(\varepsilon_{1t} - \varepsilon_{2t})\} + \varepsilon_{1t}^2 - \varepsilon_{2t}^2 + const, \\ &= -2\{-\underbrace{\xi_1 x_t^*\varepsilon_{1t}}_{I} + \underbrace{\xi_2 x_t^*\varepsilon_{2t}}_{II} + \underbrace{y_t^*\varepsilon_{1t}}_{III} - \underbrace{y_t^*\varepsilon_{2t}}_{IV}\} + \underbrace{\varepsilon_{1t}^2}_{V} - \underbrace{\varepsilon_{2t}^2}_{VI} + const, \end{aligned} \tag{20}$$

(18) becomes

$$\begin{aligned} z_t &= -2\{-x_t^*(\xi_y\widehat{y}_{2t}^* - \xi_1\eta_t) + (\varepsilon_{1t} - \widehat{y}_{2t}^*)\eta_t + \varepsilon_{1t}(\xi_y\mu_x - \mu_1)\} + \varepsilon_{1t}^2 - \widehat{y}_{2t}^{*^2} - 2\beta_y\varepsilon_{1t} - \xi_1\xi_y x_t^{*^2} + const, \\ &= -2\{-\underbrace{\xi_y x_t^*\widehat{y}_{2t}^*}_{I} + \underbrace{\xi_1 x_t^*\eta_t}_{II} + \underbrace{\varepsilon_{1t}\eta_t}_{III} - \underbrace{\widehat{y}_{2t}^*\eta_t}_{IV} + \underbrace{\varepsilon_{1t}(\xi_y\mu_x - \mu_1)}_{V}\} + \underbrace{\varepsilon_{1t}^2}_{VI} - \underbrace{\widehat{y}_{2t}^{*^2}}_{VII} - \underbrace{2\beta_y\varepsilon_{1t}}_{VIII} - \underbrace{\xi_1\xi_y x_t^{*^2}}_{IX} + const, \end{aligned} \tag{21}$$

and finally (19) is

$$\begin{aligned} z_t &= -2(\varepsilon_{1t} - \varepsilon_{2t})\eta_t + \varepsilon_{1t}^2 - \varepsilon_{2t}^2 + const, \\ &= -2\underbrace{\varepsilon_{1t}\eta_t}_{I} + \underbrace{2\varepsilon_{2t}\eta_t}_{II} + \underbrace{\varepsilon_{1t}^2}_{III} - \underbrace{\varepsilon_{2t}^2}_{IV} + const. \end{aligned} \tag{22}$$

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 (20), we have

$$I_{20} \sim \begin{cases} LM(\max\{d_x + d_{\varepsilon_1} - 1/2,\ 0\}), & \text{if } S_{x,\varepsilon_1} \neq 0 \\ LM(d_x + d_{\varepsilon_1} - 1/2), & \text{if } S_{x,\varepsilon_1} = 0 \end{cases}$$

$$II_{20} \sim \begin{cases} LM(\max\{d_x + d_{\varepsilon_2} - 1/2,\ 0\}), & \text{if } S_{x,\varepsilon_2} \neq 0 \\ LM(d_x + d_{\varepsilon_2} - 1/2), & \text{if } S_{x,\varepsilon_2} = 0 \end{cases}$$

$$III_{20} \sim \begin{cases} LM(\max\{d_y + d_{\varepsilon_1} - 1/2,\ 0\}), & \text{if } S_{y,\varepsilon_1} \neq 0 \\ LM(d_y + d_{\varepsilon_1} - 1/2), & \text{if } S_{y,\varepsilon_1} = 0 \end{cases}$$

$$IV_{20} \sim \begin{cases} LM(\max\{d_y + d_{\varepsilon_2} - 1/2,\ 0\}), & \text{if } S_{y,\varepsilon_2} \neq 0 \\ LM(d_y + d_{\varepsilon_2} - 1/2), & \text{if } S_{y,\varepsilon_2} = 0 \end{cases}$$

$$V_{20} \sim LM(\max\{2d_{\varepsilon_1} - 1/2,\ 0\})$$

$$\text{and} \qquad VI_{20} \sim LM(\max\{2d_{\varepsilon_2} - 1/2,\ 0\}).$$

Since by definition $d_x > d_{\varepsilon_i}$, the memory of V_{20} and VI_{20} is always of a lower order than that of I_{20} and II_{20}. As in the proof of Proposition 2, the squares in terms V_{20} and VI_{20} 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 (21), we have

$$I_{21} \sim \begin{cases} LM(\max\{d_x + d_2 - 1/2,\ 0\}), & \text{if } S_{x,\widehat{y}_2} \neq 0 \\ LM(d_x + d_2 - 1/2), & \text{if } S_{x,\widehat{y}_2} = 0 \end{cases}$$

$$II_{21} \sim \begin{cases} LM(\max\{d_x + d_\eta - 1/2,\ 0\}), & \text{if } S_{x,\eta} \neq 0 \\ LM(d_x + d_\eta - 1/2), & \text{if } S_{x,\eta} = 0 \end{cases}$$

$$III_{21} \sim \begin{cases} LM(\max\{d_{\varepsilon_1} + d_\eta - 1/2,\ 0\}), & \text{if } S_{\varepsilon_1,\eta} \neq 0 \\ LM(d_{\varepsilon_1} + d_\eta - 1/2), & \text{if } S_{\varepsilon_1,\eta} = 0 \end{cases}$$

$$IV_{21} \sim \begin{cases} LM(\max\{d_2 + d_\eta - 1/2,\ 0\}), & \text{if } S_{\widehat{y}_2,\eta} \neq 0 \\ LM(d_2 + d_\eta - 1/2), & \text{if } S_{\widehat{y}_2,\eta} = 0 \end{cases}$$

$$V_{21} \sim LM(d_{\varepsilon_1})$$

$$VI_{21} \sim LM(\max\{2d_{\varepsilon_1} - 1/2,\ 0\})$$

$$VII_{21} \sim LM(\max\{2d_2 - 1/2,\ 0\})$$

$$VIII_{21} \sim LM(d_{\varepsilon_1})$$

$$\text{and} \qquad IX_{21} \sim LM(\max\{2d_x - 1/2, 0\}).$$

Here, V_{21} can be disregarded since it is of the same order as $VIII_{21}$. $VIII_{21}$ dominates VI_{21}, because $d_{\varepsilon_1} < 1/2$. Finally, as $d_{\varepsilon_1} < d_x$ holds by assumption, III_{21} is dominated by II_{21} and $d_\eta < d_x$, so that IX_{21} dominates II_{21}. 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 (22), we have

$$I_{22} \sim \begin{cases} LM(\max\{d_\eta + d_{\varepsilon_1} - 1/2,\ 0\}), & \text{if } S_{\eta,\varepsilon_1} \neq 0 \\ LM(d_\eta + d_{\varepsilon_1} - 1/2), & \text{if } S_{\eta,\varepsilon_1} = 0 \end{cases}$$

$$II_{22} \sim \begin{cases} LM(\max\{d_\eta + d_{\varepsilon_2} - 1/2,\ 0\}), & \text{if } S_{\eta,\varepsilon_1} \neq 0 \\ LM(d_\eta + d_{\varepsilon_2} - 1/2), & \text{if } S_{\eta,\varepsilon_2} = 0 \end{cases}$$

$$III_{22} \sim LM(\max\{2d_{\varepsilon_1} - 1/2,\ 0\})$$

$$IV_{22} \sim LM(\max\{2d_{\varepsilon_2} - 1/2,\ 0\}).$$

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_{22} and IV_{22}. □

Proof (Proposition 5). *First note that 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. □