Identification of Volatility Proxies as Expectations of Squared Financial Return

Sucarrat, Genaro

BI Norwegian Business School

20 July 2020

Identification of Volatility Proxies as Expectations of Squared Financial Return*

Genaro Sucarrat†

20th July 2020

Abstract

Volatility proxies like Realised Volatility (RV) are extensively used to assess the forecasts of squared financial return produced by Autoregressive Conditional Heteroscedasticity (ARCH) models. But are volatility proxies identified as expectations of the squared return? If not, then the results of these comparisons can be misleading, even if the proxy is unbiased. Here, a tripartite distinction between strong, semi-strong and weak identification of a volatility proxy as an expectation of squared return is introduced. The definition implies that semi-strong and weak identification can be studied and corrected for via a multiplicative transformation. Well-known tests can be used to check for identification and bias, and Monte Carlo simulations show they are well-sized and powerful – even in fairly small samples. As an illustration, twelve volatility proxies used in three seminal studies are revisited. Half of the proxies do not satisfy either semi-strong or weak identification, but their corrected transformations do. Correcting for identification does not always reduce the bias of the proxy, so there is a tradeoff between the choice of correction and the resulting bias.

Keywords: GARCH models, financial time-series econometrics, volatility forecasting, Realised Volatility

1 Introduction

Let $\{r_t^2\}$ denote a discrete time process of squared financial returns defined on the probability space $(\Omega, \mathcal{F}, P)$. Often, r_t^2 can be expressed as

$$r_t^2 = \sigma_t^2 \eta_t^2, \tag{1}$$

where $\sigma_t^2 > 0$ *a.s.* is a scale or volatility and $\eta_t^2 \geq 0$ *a.s.* is an innovation. The decomposition is not unique, since many pairs $\{\sigma_t^2\}$ and $\{\eta_t^2\}$ may satisfy (1). Clearly, for a comparison between two different models σ_{1t}^2 and σ_{2t}^2 to be meaningful, they must be on the same scale. For example, if the former corresponds to the conditional variance while the target of the latter is the double

*I am grateful to Steffen Grønneberg, David Kreiberg, Sebastien Laurent and Giuseppe Storti for useful comments and suggestions.

†Department of Economics, BI Norwegian Business School, Nydalsveien 37, 0484 Oslo, Norway. Email genaro.sucarrat@bi.no, phone +47+46410779. Webpage: http://www.sucarrat.net/

of that, then one or the other must be adjusted before comparison. Another possibility is that σ_{2t}^2 measures σ_{1t}^2 with error, say, $\sigma_{2t}^2 = \sigma_{1t}^2 \epsilon_t$, where $\epsilon_t \geq a.s.$ is the measurement error. Even if the properties of ϵ_t are such that the expectation of σ_{2t}^2 is equal to σ_{1t}^2, the presence of the measurement error ϵ_t may change the scale of σ_{2t}^2. Again, if this is the case, then one or the other must be adjusted before comparison.

The assumed or entertained scale σ_t^2 is unobserved, and this creates a challenge in *ex post* forecast evaluation. One solution that has been put forward is to use high-frequency intraperiod financial data to construct an observable volatility proxy

$$V_t > 0 \quad a.s.$$

for σ_t^2, and then to evaluate an estimate $\widehat{\sigma}_t^2$ against V_t. See, for example, Park and Linton (2012), and Violante and Laurent (2012) for surveys of this approach. Realised Volatility (RV), i.e. the sum of intraperiod squared returns, is the most commonly used volatility proxy, and a popular metric of forecast precision within this approach is the Mean Squared Error (MSE): $T^{-1}\sum_{t=1}^{T}(V_t - \widehat{\sigma}_t^2)^2$. Subject to suitable assumptions, the volatility proxy V_t in question tends to a limit σ_{Vt}^2 as the intraperiod sampling frequency increases towards infinity. For RV, the limit σ_{Vt}^2 is the Integrated Variance (IV), which may – or may not – be equal to the assumed or entertained specification σ_t^2. While σ_{Vt}^2 may differ from σ_t^2 even for simple specifications of σ_t^2, e.g. the first order Generalised ARCH (GARCH), it is particularly likely to happen in explanatory modelling of financial variability, where additional covariates are considered as predictors and/or explanatory variables in the specification of σ_t^2, see Sucarrat (2009) for a discussion. Another complication is that, in empirical practice, the sampling frequency is finite, and the observations used to compute the volatility proxy V_t are often contaminated by market microstructure noise. So it is widely believed that V_t measures σ_{Vt}^2 with error, e.g. multiplicatively, $V_t = \sigma_{Vt}^2 \epsilon_t$, or additively, $V_t = \sigma_{Vt}^2 + \epsilon_t$. See e.g. Andersen et al. (2005), Bandi and Russell (2008), Aït-Sahalia and Mykland (2009), Bollerslev et al. (2016), Yeh and Wang (2019), and the numerous references therein. In spite of the measurement error ϵ_t and the possibility that σ_{Vt}^2 may not equal the entertained specification of σ_t^2, there is a widespread belief that a suitably computed proxy V_t may provide an efficient – but not necessarily unbiased – estimate of the entertained specification of σ_t^2. This is why many studies use a volatility proxy as a substitute for the assumed specification of σ_t^2, and evaluate volatility forecasts $\{\widehat{\sigma}_t^2\}$ against $\{V_t\}$.

Arguably, the most common specifications of σ_t^2 belong to the Autoregressive Conditional Heteroscedasticity (ARCH) class of models proposed by Engle (1982). In that case, σ_t^2 corresponds to the conditional expectation of r_t^2. A volatility model σ_t^2 is equal to the expectation of r_t^2 conditional on a σ-field $\mathcal{F}_{t-1} \subset \mathcal{F}$ if

$$\sigma_t^2 = E(r_t^2 | \mathcal{F}_{t-1}).$$

If this holds, then two main properties follow under stationarity:

$$\begin{array}{lll} \text{Unbiasedness:} & E(r_t^2 - \sigma_t^2 | \mathcal{F}_{t-1}) = 0 \quad \text{and} \quad E(r_t^2 - \sigma_t^2) = 0, \\ \text{Identification:} & E(r_t^2/\sigma_t^2 | \mathcal{F}_{t-1}) = 1 \quad \text{and} \quad E(r_t^2/\sigma_t^2) = 1. \end{array}$$

It is the second of these properties that is the primary focus of this paper. Define

$$\eta_t^2 := r_t^2/\sigma_t^2,$$

where σ_t^2 is a model of r_t^2. Borrowing from the terminology of Drost and Nijman (1993), a specification σ_t^2 is said to be strongly, semi-strongly or weakly[1] identified as an expectation of r_t^2 if:

$$\begin{aligned} \text{Strong identification:} \quad & \eta_t^2 \sim iid \text{ with } E(\eta_t^2) = 1 \text{ for all } t, && (2) \\ \text{Semi-strong identification:} \quad & E(\eta_t^2|\mathcal{F}_{t-1}) = 1, \quad \mathcal{F}_{t-1} \subset \mathcal{F}, \quad \text{for all } t, && (3) \\ \text{Weak identification:} \quad & E(\eta_t^2) = 1 \quad \text{for all } t. && (4) \end{aligned}$$

Note that, in (3), identification is with respect to a σ-field $\mathcal{F}_{t-1}$. Of course, (2) $\Rightarrow$ (3) and (3) $\Rightarrow$ (4), but their converses are not true. ARCH models are examples of σ_t^2 for which one or more of these definitions usually hold, whereas Stochastic Volatility (SV) models are examples for which one or more of the definitions usually do not hold. A model σ_t^2 for which weak identification always hold is $\sigma_t^2 = E(r_t^2)$.

Suppose σ_t^2 is a model of r_t^2 that is either strongly, semi-strongly or weakly identified as an expectation of r_t^2. For a volatility proxy V_t to be a valid proxy for σ_t^2, it should satisfy identifiability criteria similar to (2)–(4). Otherwise, V_t is not at the same scale-level as σ_t^2. For SV models, by contrast, where σ_t^2 is not an expectation of r_t^2, it is not clear that similar identifiability criteria should be required. Define

$$z_t^2 := r_t^2/V_t.$$

The volatility proxy V_t is strongly, semi-strongly or weakly identified as an expectation of r_t^2 if:

$$\begin{aligned} \text{Strong identification:} \quad & z_t^2 \sim iid \text{ with } E(z_t^2) = 1 \text{ for all } t, && (5) \\ \text{Semi-strong identification:} \quad & E(z_t^2|\mathcal{F}_{t-1}) = 1, \quad \mathcal{F}_{t-1} \subset \mathcal{F}, \quad \text{for all } t, && (6) \\ \text{Weak identification:} \quad & E(z_t^2) = 1 \quad \text{for all } t. && (7) \end{aligned}$$

Again, semi-strong identification is with respect to a σ-field $\mathcal{F}_{t-1}$, and again (5) $\Rightarrow$ (6) and (6) $\Rightarrow$ (7). Some useful properties follow directly from (5)–(7). First, if $h_t := E(z_t^2|\mathcal{F}_{t-1})$ exists for all t, then a volatility proxy V_t can be transformed to satisfy semi-strong identification via a multiplicative transformation:

$$h_t V_t \qquad \text{satisifies} \quad E\left(r_t^2/(h_t V_t)|\mathcal{F}_{t-1}\right) = 1 \quad \text{for all } t. \tag{8}$$

In particular, if $h := E(z_t^2)$ exists for all t, then a volatility proxy V_t can always be transformed to satisfy weak identification:

$$hV_t \quad \text{satisfies} \quad E\left(r_t^2/(hV_t)\right) = 1 \quad \text{for all } t. \tag{9}$$

Practical procedures for identification are thus widely available in public software: The sample average $T^{-1}\sum_{t=1}^{T} z_t^2$ provides a consistent estimate of h subject to fairly mild assumptions, and Multiplicative Error Models (MEMs) naturally suggest themselves as models of h_t, see Brownlees et al. (2012) for a survey of MEMs.[2] These considerations suggest the following procedure whenever an observed volatility proxy V_t is considered as a substitute for an expectation σ_t^2 of r_t^2:

[1] While the terms "strong" and "semi-strong" are used in similar ways to Drost and Nijman (1993), the way the term "weak" is used differs.

[2] MEMs are essentially GARCH-models of non-negative variables. This was first noted by Engle and Russell (1998).

1. Check whether the proxy V_t is identified as an expectation. That is, check whether it satisfies one or more of the criteria in (5)–(7).

2. If V_t is not identified according to Step 1, choose a suitable specification h_t to contruct an identification corrected proxy $h_t V_t$. To this end, attention should be paid to how the choice of h_t affects the bias of $h_t V_t$ for r_t^2. Since unbiasedness and identification are not equivalent, there might be a trade-off between the choice of h_t and the magnitude of the bias. Some choices of h_t may reduce the bias, others may increase it.

3. Compare estimates $\{\widehat{\sigma}_t^2\}$ against the identification corrected proxy $\{\widehat{h}_t V_t\}$ rather than against V_t.

In the empirical illustration of this procedure in Section 5, the focus is on steps 1 and 2.

This paper makes five contributions. First, the tripartite distinction between strong, semi-strong and weak identification of a volatility proxy as an expectation of squared return is introduced. This was done above in (5)–(7). The multiplicative transformation involved in the definition of identification implies that a volatility proxy can be corrected to satisfy identification in a straightforward manner, recall (8) and (9), and leads to the three-step procedure outlined above. Second, a set of well-known tests that can be used to check a volatility proxy for semi-strong and weak identification is proposed and evaluated. Arguably, semi-strong and weak identification are of greater interest than strong identification, since the independence and identicality assumptions associated with strong identification will often not hold in practice. The focus is on tests that are readily implemented in widely available software, and Monte Carlo simulations show the tests are well-sized and powerful, even in fairly small samples. In a third contribution the specification of h_t is disucssed. While MEMs naturally suggest themselves, it is shown that, under strict stationarity and ergodicity of $\{z_t^2\}$, the process admits a representation that is particularly useful. Specifically, it is shown that $\{z_t^2\}$ admits a log-MEM(p,0) representation – i.e. a MEM of the log-ARCH type – whose parameters can straightforwardly be estimated consistently by means of a least squares procedure. The log-MEM specification is of special interest, since our empirical illustration reveals z_t^2 is often negatively autocorrelated (MEMs of the ARCH type are not compatible with negative autocorrelations). A fourth contribution consists of shedding new light on tests for bias via regressions of the Mincer and Zarnowitz (1969) (MZ) type. It is shown that, in general, the Standard MZ-test is flawed when V_t measures σ_t^2 with error: The null of no bias is erroneously rejected with probability 1 as $T \to \infty$. However, straightforward modifications to the test rectifies the flaw. Monte Carlo simulations show that the simplest of the modifications is particularly well-sized – even in small samples, since the discrepancy between the empirical and nominal sizes is less than 1%-point already for $T = 500$ in the simulations. In a fifth contribution, an empirical illustration, twelve volatility proxies used in three seminal studies are revisited. Out of the twelve proxies, half of them are found to either not satisfy weak or semi-strong identification, or both. Next, estimates of h_t are used to construct corrected proxies that satisfy either weak or semi-strong identification, or both. Interestingly, z_t^2 is usually negatively autocorrelated, which means MEMs of the non-exponential ARCH type are not appropriate as models of h_t for the investigated proxies. Instead, a log-MEM(1,0) – i.e. a MEM of the log-ARCH(1) type – is found to be a suitable specification of h_t in most of the cases. Identification correction does not always lead to a reduction in bias, thus illustrating the tradeoff between the chosen specification of h_t and the resulting bias.

The rest of the paper is organised as follows. The next section, Section 2, contains the proposed tests for identification, together with Monte Carlo simulations of their size and power. Section 3 discusses the specification of h_t, and contains the result on the existence of a log-MEM(p,0) representation of $\{z_t^2\}$. In Section 4 tests of the MZ-type for bias are revisited. Section 5 contains the empirical illustration, whereas Section 6 concludes.

2 Tests for identification

The focus is on tests that are easy to implement, widely available and well-sized without the need for size-correction. Four tests are proposed. The first two are based on the sample average, and can be used to test whether h differs from 1, i.e. whether a volatility proxy is weakly identified or not. The next two test for autocorrelation in z_t^2 and $\ln z_t^2$, respectively, and can thus be used to test for departures from semi-strong identification. The section ends by studying the finite sample size and power of the tests via Monte Carlo simulations.

2.1 Tests based on the sample average

Subject to fairly mild assumptions, the sample average $\widehat{h} = T^{-1}\sum_{t=1}^{T} z_t^2$ provides a consistent estimate of $E(z_t^2) = h$. Strong, semi-strong and weak identification all require that $h = 1$. Since $\widehat{h}$ is also the Least Squares (LS) estimate of h in the linear regression $z_t^2 = h + u_t$, we can readily implement tests of $h = 1$ with widely available software when u_t is heteroscedastic or autocorrelated, or both. Specifically, if

$$\sqrt{T}(\widehat{h} - h) \sim N(0, \Sigma) \tag{10}$$

asymptotically and there exists a consistent estimator $\widehat{\Sigma}$ for Σ, then the test can be implemented as

$$\text{Test 1:} \qquad \frac{\widehat{h} - 1}{se(\widehat{h})} \sim t(T-1), \qquad H_0 : h = 1 \qquad vs. \qquad H_A : h \neq 1, \tag{11}$$

where $se(\widehat{h}) = (\widehat{\Sigma}/T)^{1/2}$ is the standard error of $\widehat{h}$ returned by the software. The option to select either an ordinary, heteroscedasticity robust or Heteroscedasticity and Auto-Correlation (HAC) robust standard error is widely available. Often, the latter two are those of White (1980), Newey and West (1987), respectively. If strong identification holds, then u_t is *iid*, and so the ordinary standard error is suitable. Under semi-strong identification, however, the u_t's can be heteroscedastic. If this is the case, then a heteroscedasticity robust standard error is more suitable. Under weak identification, z_t^2 can also be autocorrelated. If this is the case, then a HAC robust standard error is more suitable. Below, in the simulations, the size and power for the HAC robust standard error of Newey and West (1987) is investigated. As we will see, the empirical size corresponds well to the nominal size.

The distribution of z_t^2 will usually have an exponential-like shape, so tests based on the average of $\ln z_t^2$ may be more efficient. The results in Sucarrat et al. (2016) can be used to build a regression-like test, where $\widehat{\phi} = T^{-1}\sum_{t=1}\ln z_t^2$ estimates ϕ in $\ln z_t^2 = \phi + u_t$ in a first step, and then the residuals are used in a second step to complete an estimate of $\ln h$. Interestingly, this two-step estimator is numerically identical to[3]

$$\ln \widehat{h}$$

when there are no zeros in $\{z_t^2\}$. In other words, if (10) holds, then the delta method straightforwardly leads to

$$\sqrt{T}(\ln \widehat{h} - \ln h) \sim N(0, \Sigma/h^2),$$

where Σ is the same asymptotic variance as in (10). This means the asymptotic variance of $\ln \widehat{h}$ is smaller (greater) than that of $\widehat{h}$ when $h > 1$ ($h < 1$). Below, in the simulations, the test is implemented as

$$\text{Test 2:} \qquad \frac{\ln \widehat{h}}{se(\widehat{h})/\widehat{h}} \sim t(T-1), \qquad H_0 : \ln h = 0 \qquad vs. \qquad H_A : \ln h \neq 0, \tag{12}$$

where $se(\widehat{h}) = (\widehat{\Sigma}/T)^{1/2}$ is the standard error of Newey and West (1987). As we will see, the test in (12) is indeed more (less) powerful than (11) in finite samples when $h > 1$ ($h < 1$).

2.2 Tests for autocorrelation

If semi-strong identification holds, then $\{z_t^2\}$ is not autocorrelated. Tests for autocorrelation in z_t^2 can therefore be used to test whether semi-strong identification holds or not. Additionally, tests for autocorrelation in z_t^2 can also be used to shed light on whether h_t is suitably modelled as a MEM or log-MEM. Because if h_t is a stationary MEM(p, q) of the GARCH type, then z_t^2 will have positive autocorrelations, see Francq and Zakoïan (2019, p. 47). In other words, if negative autocorrelations are present, then h_t is more suitably modelled as a log-MEM.

A well-known and widely available test for autocorrelation that suggests itself is the Portmanteau test of Ljung and Box (1979). Its test statistic for autocorrelation up to and including order p is given by

$$\text{Test 3:} \qquad T(T+2)\sum_{i=1}^{p}\frac{\widehat{\rho}_i(z_t^2)}{(T-i)} \sim \chi^2(p), \tag{13}$$

where $\widehat{\rho}_i(z_t^2)$ is the sample correlation between z_t^2 and z_{t-i}^2. Note that, asymptotically, this test is in fact equivalent to an LM-test of h_t being a MEM$(p, 0)$ with $p = 0$ under the null, see Francq and Zakoïan (2019, pp. 147-148). Below, in the simulations, the size and power of H_0: $Corr(z_t^2, z_{t-1}^2) = 0$ and H_A: $Corr(z_t^2, z_{t-1}^2) \neq 0$, respectively, is studied.

Another possibility is that $\ln z_t^2$ is autocorrelated. This is the case, for example, if $\ln h_t$ is a stationary log-MEM of the log-GARCH type. In this case $\ln z_t^2$ admits an ARMA(p, q) representation,[4] and so $\ln z_t^2$ will be autocorrelated under the usual ARMA-conditions, see

[3]When there are no zeros in $\{z_t^2\}$, the sample average $\widehat{\phi} = T^{-1}\sum_{t=1}^{T}\ln z_t^2$ provides an estimate of ϕ in the regression $\ln z_t^2 = \phi + u_t$. The second-step estimator implied by Sucarrat et al. (2016) is $\widehat{\tau} = \ln T^{-1}\sum_{t=1}^{T} e^{\widehat{u}_t}$ with $\widehat{u}_t = \ln z_t^2 - \widehat{\phi}$. Combining them gives $\widehat{\phi} + \widehat{\tau} = \ln \widehat{h}$.

[4]The existence of the ARMA representation requires that the zero-probability is zero so that $E|\ln z_t^2| < \infty$. This usually holds for return series of liquid stocks, for which volatility proxies based on intraday data are usually considered.

Sucarrat (2019). Also here the Portmanteau test of Ljung and Box (1979) is a natural candidate. The test statistic in this case is

$$\text{Test 4:} \qquad T(T+2)\sum_{i=1}^{p}\frac{\widehat{\rho}_i(\ln z_t^2)}{(T-i)} \sim \chi^2(p), \tag{14}$$

where $\widehat{\rho}_i(\ln z_t^2)$ is now the sample correlation between $\ln z_t^2$ and $\ln z_{t-i}^2$. Below, in the simulations, the size and power of H_0: $Corr(\ln z_t^2, \ln z_{t-1}^2) = 0$ and H_A: $Corr(\ln z_t^2, \ln z_{t-1}^2) \neq 0$, respectively, is studied.

2.3 Monte Carlo simulations

In this subsection the size and power of four tests are studied:

	H_0	H_A	Test statistic
Test 1:	$h = 1$	$h \neq 1$	(11)
Test 2:	$\ln h = 0$	$\ln h \neq 0$	(12)
Test 3:	$Corr(z_t^2, z_{t-1}^2) = 0$	$Corr(z_t^2, z_{t-1}^2) \neq 0$	(13) with $p = 1$
Test 4:	$Corr(\ln z_t^2, \ln z_{t-1}^2) = 0$	$Corr(\ln z_t^2, \ln z_{t-1}^2) \neq 0$	(14) with $p = 1$

Two classes of Data Generating Processes (DGPs) are used in the experiments:

$$\begin{array}{lrcl}
 & z_t^2 & = & h_t\eta_t^2, \qquad \eta_t \overset{iid}{\sim} N(0,1), \qquad t = 1,\ldots,T, \\
\text{DGP 1:} & h_t & = & h, \qquad h \in \{0.9, 1, 1.1\}, \qquad E(z_t^2) = h, \\
\text{DGP 2:} & \ln h_t & = & \omega + \alpha \ln z_{t-1}^2, \qquad \boldsymbol{\theta} = (\omega, \alpha)', \\
\quad\text{a)} & \boldsymbol{\theta} & = & (-0.16, -0.1)', \qquad E(z_t^2) = 1.00, \qquad Corr(z_t^2, z_{t-1}^2) = -0.0.9, \\
\quad\text{b)} & \boldsymbol{\theta} & = & (0, -0.1)', \qquad E(z_t^2) = 1.15, \qquad Corr(z_t^2, z_{t-1}^2) = -0.0.9, \\
\quad\text{c)} & \boldsymbol{\theta} & = & (0, 0.1)', \qquad E(z_t^2) = 0.89, \qquad Corr(z_t^2, z_{t-1}^2) = 0.10,
\end{array}$$

In the first class, $\{z_t^2\}$ is iid with $E(z_t^2) = h$. So strong identification holds when $h = 1$, and all three kinds of identification fail when $h \neq 1$. In the second class, the DGP is a log-MEM of the log-ARCH(1) type. The choice of specification is informed by the empirical results in Section 5. In 2a), $E(z_t^2) = 1$ and $Corr(z_t^2, z_{t-1}^2) = -0.0.9$, so weak identification holds but not semi-strong identification. In 2b) and 2c) both semi-strong and weak identification fail.

Table 1 contains the simulation results of Tests 1 and 2. In these tests the null $E(z_t^2) = 1$ holds in two experiments: DGP 1 with $h_t = 1$ and DGP 2a). For these experiments, the empirical rejection frequencies correspond well to their nominal levels (10%, 5% and 1%). Indeed, the empirical levels are never more than 1.3 percentage-points away from their nominal counterparts. Turning to the power of the tests, the alternative hypothesis $E(z_t^2) \neq 1$ holds in four experiments: DGP 1 with $h_t = 1.1$, DGP 1 with $h_t = 0.9$, DGP 2b) and DGP 2c). The results show that the tests are very powerful in sample sizes of practical relevance. For $T = 5000$, for example, which is fairly common in empirical work, the probability of rejecting is greater than 98% in all three experiments. For smaller sample sizes, the results show that the tests have notable power already at $T = 250$, which is an unusually low sample size in empirical work. As for relative power, Test 1 is more powerful than Test 2 when $E(z_t^2) = h < 1$, and the opposite is the case when $E(z_t^2) = h > 1$. This is in line with the expression of the asymptotic variance of Test 2. The results show that the difference in power is larger the smaller the sample size T.

Table 2 contains the simulation results of Tests 3 and 4. In these tests the null, $Corr(z_t^2, z_{t-1}^2) = 0$ or $Corr(\ln z_t^2, \ln z_{t-1}^2) = 0$, holds in the DGP 1 experiment where $h_t = 1$ for all t. Again, the empirical rejection frequencies correspond well to their nominal levels (10%, 5% and 1%) under the null, since the empirical levels are never more than 1 percentage-point away from their nominal counterparts. The alternative hypotheses of Tests 3 and 4 hold in three experiments: DGP 2a), DGP 2b) and DGP 2c). Again the results show that the tests are very powerful in sample sizes of practical relevance. Already at $T = 2000$ the rejection frequency is 93% or higher for a 1% significance level. For $T = 5000$, which is fairly common in empirical work, the probability of rejecting is greater than 98% in all three experiments. For smaller sample sizes, the results show that the tests have notable power already for $T = 250$, which is an unusually low sample size in empirical work. As for a comparison of power, Test 4 is usually more powerful than Test 3. This is particularly the case in small samples, i.e. $T = 250$ and $T = 500$. As the sample size grows, however, the results are more mixed.

3 Specification of h_t

If z_t^2 is ergodic stationary and $E|z_t^2| < \infty$, then $h = E(z_t^2)$ is consistently estimated by the sample average. For time-varying specifications of h_t, there is a wide range of alternatives available. In particular, Multiplicative Error Models (MEMs) suggest themselves as models of h_t, see Brownlees et al. (2012) for a survey of MEMs.

The MEM counterpart of the GARCH(p, q) model is

$$z_t^2 = h_t u_t, \qquad E(u_t|\mathcal{F}_{t-1}) = 1 \qquad \text{for all } t, \tag{15}$$

$$h_t = \omega + \sum_{i=1}^{p} \alpha_i z_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}, \qquad \omega > 0, \qquad \alpha_i, \beta_j \geq 0. \tag{16}$$

Unfortunately, this subclass of MEMs is not compatible with negative autocorrelations on z_t^2, see Proposition 2.2 in Francq and Zakoïan (2019, p. 47). And, as we will see in Section 5, negative autocorrelations are common empirically. Log-MEMs, by contrast, are compatible with negative autocorrelations on z_t^2. Define

$$y_t = \begin{cases} \ln z_t^2 & \text{if } z_t^2 \neq 0 \\ 0 & \text{if } z_t^2 = 0 \end{cases}. \tag{17}$$

The zero-augmented log-MEM(p, q) is given by (15) together with

$$\ln h_t = \omega + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \beta_j \ln h_{t-j}. \tag{18}$$

Note that there are no non-negativity restrictions on the parameters. While $z_t^2 = 0$ is unlikely in returns for which high-frequency intraperiod data is available, there is no loss of generality in allowing for zeros by defining y_t as in (17). A variant of (18) was proposed by Hautsch et al. (2013) for volume, and the extended log-GARCH of Francq and Zakoïan (2019, Section 4.3) nests (18) as a special case.

A subclass of log-MEMs that is of special interest in the current context is the log-MEM$(p, 0)$, i.e. $\ln h_t = \boldsymbol{x}_t'\boldsymbol{b}$, where $\boldsymbol{x}_t = (1, y_{t-1}, \ldots, y_{t-p})'$ and $\boldsymbol{b} = (\omega, \alpha_1, \ldots, \alpha_p)'$. The reason is that, subject to fairly general and mild assumptions, z_t^2 admits a weak log-MEM$(p, 0)$ representation regardless of whether the DGP is a log-MEM or not, see Proposition 1 below. The result relies on assumptions that ensures the Ordinary Least Squares (OLS) estimator

$$\widehat{\boldsymbol{b}}_T^* = \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t y_t\right)$$

converges to a limit $\boldsymbol{b}^* = (\omega^*, \alpha_1, \ldots, \alpha_p)'$. Next, define

$$\ln h_t^* := \omega^* + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p}$$

and

$$\ln h_t := \omega + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p}, \qquad \omega := \omega^* + \ln E(u_t^*), \qquad u_t^* := z_t^2/h_t^*. \tag{19}$$

By construction,

$$z_t^2 = h_t^* u_t^* = h_t u_t \qquad \text{with} \qquad E(u_t) = 1,$$

which means (19) is a weak log-MEM$(p, 0)$ representation of z_t^2. Subject to suitable assumptions,

$$\widehat{E}(u_t^*) = \frac{1}{T}\sum_{t=1}^T \widehat{u}_t^*, \qquad \widehat{u}_t^* = \frac{z_t^2}{\exp(\boldsymbol{x}_t'\widehat{\boldsymbol{b}}_T^*)}, \tag{20}$$

is consistent for $E(u_t^*)$, and $\widehat{\omega} = \widehat{\omega}^* + \ln \widehat{E}(u_t^*)$ is consistent for ω. Note that (20) is simply the smearing estimator of Duan (1983). If, in addition, $E(u_t|\mathcal{F}_{t-1}) = 1$ for all t, then it follows straightforwardly that $h_t V_t$ satisfies semi-strong identification.

Proposition 1 *Suppose $\{z_t^2\}$ and $\{y_t\}$ are ergodic stationary and measurable, $E(\boldsymbol{x}_t\boldsymbol{x}_t')$ is finite and nonsingular for all t, and $E|u_t^*| < \infty$ and $\widehat{E}(u_t^*) \overset{a.s.}{\to} E(u_t^*)$. Then there exists a representation*

$$z_t^2 = h_t u_t, \qquad \ln h_t = \omega + \sum_{i=1}^p \alpha_i y_{t-i}, \qquad E(u_t) = 1, \tag{21}$$

with $\widehat{\boldsymbol{b}}_T \overset{a.s.}{\to} \boldsymbol{b}$, where $\widehat{\boldsymbol{b}}_T = (\widehat{\omega}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_p)'$ and $\boldsymbol{b} = (\omega, \alpha_1, \ldots, \alpha_p)'$. If, in addition, $E(u_t|\mathcal{F}_{t-1}) = 1$ for all t, then $h_t V_t$ satisfies semi-strong identification.

Proof: The ergodic stationarity and measurability of $\{z_t^2\}$ and $\{y_t\}$ means each entry in $\boldsymbol{x}_t\boldsymbol{x}_t'$ and $\boldsymbol{x}_t y_t$ is ergodic stationary. Accordingly, by the ergodic theorem, the finiteness and nonsingularity of $E(\boldsymbol{x}_t\boldsymbol{x}_t')$, and the continuous mapping theorem, the OLS estimator $\widehat{\boldsymbol{b}}_T^*$ converges almost surely to a limit $\boldsymbol{b}^*$. Next, the assumption $\widehat{E}(u_t^*) \overset{a.s.}{\to} E(u_t^*)$ implies $\widehat{\boldsymbol{b}}_T \overset{a.s.}{\to} \boldsymbol{b}$. Finally, semi-strong identification follows directly if $E(u_t|\mathcal{F}_{t-1}) = 1$ for each t. $\square$

A similar result can be derived for MEMs of the ARCH(p) type. However, that result is less interesting, since it is not valid in the presence of negative autocorrelations on z_t^2. The existence of the weak log-MEM($p, 0$) representation relies on assumptions that are very mild. So existence is likely to hold in a vast range of situations. The assumption $E(u_t|\mathcal{F}_{t-1}) = 1$ for all t is less mild. If it does hold, then u_t is not autocorrelated. In empirical practice, therefore, checking whether the residuals $\widehat{u}_t$'s are autocorrelated or not can be useful in the search for a suitable order p. If $z_t^2 \neq 0$ $a.s.$, then $\boldsymbol{b}_T^*$ equals the LS estimator of the AR(p) representation $\ln z_t^2 = \ln h_t^* + \ln u_t^*$, where $E(\ln u_t^*) = 0$, see Sucarrat et al. (2016). In other words, in this case widely available software can be used to test whether one or more of the slope coefficients $\alpha_1, \ldots, \alpha_p$ are different from zero or not. For example, if $\ln u_t^*$ is heteroscedastic or autocorrelated, or both, then robust coefficient-covariance is usually available in widely available public software. Finally, note that the specification of $\ln h_t$ in (19) can straightforwardly be augmented with stochastic conditioning covariates. Minor changes to Proposition 1 and its proof would be required.

4 Tests for bias

It is possible for a proxy V_t to be identified but biased, and vise versa it is possible for a proxy V_t to be unbiased but not identified. In empirical practice, therefore, unless V_t measures σ_t^2 with no error (i.e. $\sigma_t^2 = V_t$ $a.s.$), identification correction may either reduce or increase the bias. This necessitates estimates and tests for bias. A volatility proxy V_t is conditionally or unconditionally unbiased for σ_t^2 and $E(\sigma_t^2)$, respectively, if

$$\text{Conditional unbiasedness:} \quad E(V_t|\mathcal{F}_{t-1}) = \sigma_t^2 \;\; a.s. \quad \text{for all } t, \tag{22}$$

$$\text{Unconditional unbiasedness:} \quad E(V_t) = E(\sigma_t^2) \quad \text{for all } t. \tag{23}$$

Of course, the former implies the latter, but the latter does not imply the former. Estimation and testing of conditional unbiasedness is, in general, infeasible, since σ_t^2 is unobserved. Estimation and testing of unconditional unbiasedness, however, is feasible.

4.1 Tests via Mincer-Zarnowitz regressions

Under ergodic stationarity of $\{r_t^2\}$ and $\{V_t\}$, and if $E(r_t^2) = E(\sigma_t^2)$ as in the ARCH-class of models, the sample average $T^{-1}\sum_{t=1}^{T}(r_t^2 - V_t)$ provides a consistent estimate of the unconditional bias $E(\sigma_t^2 - V_t)$. This property is exploited in tests implemented via Mincer and Zarnowitz (1969) regressions:

$$r_t^2 = \phi_0 + \phi_1 V_t + w_t.$$

Usually, ϕ_0 and ϕ_1 are estimated by OLS, and the Standard MZ-test is implemented as

$$\text{Standard MZ-test:} \quad H_0: \phi_0 = 0 \cap \phi_1 = 1 \quad vs. \quad H_A: \phi_0 \neq 0 \cup \phi_1 \neq 1, \quad W \sim \chi^2(2), \tag{24}$$

where W is the Wald-statistic. Below, in the simulations, the heteroscedasticity and autocorrelation robust coefficient-covariance of Newey and West (1987) is used to compute the Wald-statistic of this test.

If V_t measures σ_t^2 with error, then the Standard MZ-test above is flawed. The reason is that, in general, the Standard MZ-test will reject H_0 with probability 1 as $T \to \infty$, even if $E(\sigma_t^2) = E(V_t)$. To see this, consider first the case where $\sigma_t^2 = V_t$ $a.s.$, i.e. the case where there si no measurement error. The population values of ϕ_1 and ϕ_0 are then equal to those postulated by the null hypothesis: $\phi_1 = Cov(r_t^2, V_t)/Var(V_t) = 1$ and $\phi_0 = E(r_t^2) - \phi_1 E(V_t) = 0$, since

$$E(r_t^2) = E(V_t) \qquad \text{and} \qquad Cov(r_t^2, V_t) = Cov(\sigma_t^2, V_t) = Var(V_t).$$

If, instead, V_t measures σ_t^2 with error so that V_t is *not* equal to σ_t^2 $a.s.$, then we will in general have

$$Cov(r_t^2, V_t) \neq Cov(\sigma_t^2, V_t) \neq Var(V_t).$$

As a consequence, $\phi_1 \neq 1$ and $\phi_0 \neq 0$, in general. In fact, under strict stationarity and ergodicity of $\{r_t^2\}$ and $\{V_t\}$, and if $E(r_t^2) = E(V_t)$, we have

$$\phi_1 = Cov(r_t^2, V_t)/Var(V_t), \qquad \phi_0 = (1 - \phi_1)E(r_t^2) \qquad \Leftrightarrow \qquad \phi_0 + \phi_1 = 1.$$

This leads to the Modified MZ-test:

$$\text{Modified MZ-test:} \quad H_0 : \phi_0 + \phi_1 = 1 \quad vs. \quad H_A : \phi_0 + \phi_1 \neq 1, \quad W \sim \chi^2(1), \tag{25}$$

where W is the associated Wald-statistic. Below, in the simulations, the coefficient-covariance of Newey and West (1987) is used to compute the statistic. As we will see, the simulations confirm that the test rectifies the flaw of the Standard MZ-test in the presence of measurement error. However, the simulations also reveal that the Modified MZ-test is poorly sized in small and medium sized samples.

A restricted version of the MZ-test both rectifies the flaw of the Standard MZ-test, and is well-sized across small, medium and large samples. Under the null of unconditional unbiasedness, we have

$$(r_t^2 - V_t) = \phi_0 + w_t \qquad \text{with} \qquad \phi_0 = 0.$$

This leads to the Restricted MZ-test:

$$\text{Restricted MZ-test:} \qquad H_0 : \phi_0 = 0 \quad v.s. \quad H_A : \phi_0 \neq 0, \qquad \frac{\widehat{\phi}_0}{se(\widehat{\phi}_0)} \sim t(T-1), \tag{26}$$

where $\widehat{\phi}_0$ is the sample average of $(r_t^2 - V_t)$. Below, in the simulation, $se(\widehat{\phi}_0)$ is the standard error of Newey and West (1987).

4.2 Monte Carlo simulations

In this subsection the empirical size of the three tests are studied:

	H_0	H_A	Test statistic
Standard MZ-test:	$\phi_0 = 0 \ \cap \ \phi_1 = 1$	$\phi_0 \neq 0 \ \cup \ \phi_1 \neq 1$	(24)
Modified MZ-test:	$\phi_0 + \phi_1 = 1$	$\phi_0 + \phi_1 \neq 1$	(25)
Restricted MZ-test:	$\phi_0 = 0$	$\phi_0 \neq 0$	(26)

In the simulations the true volatility process $\{\sigma_t^2\}$ is governed by the GARCH(1,1) model

$$r_t^2 = \sigma_t^2 \eta_t^2, \qquad \eta_t \overset{iid}{\sim} N(0,1), \qquad \sigma_t^2 = 0.2 + 0.1 r_{t-1}^2 + 0.8 \sigma_{t-1}^2,$$

and the volatility proxy V_t is linked to σ_t^2 by

$$V_t = \sigma_t^2 \epsilon_t, \qquad \{\sigma_t^2\} \perp \{\epsilon_t\}, \qquad E(\epsilon_t) = 1, \qquad \epsilon_t = E(\varepsilon_t)^{-1} \varepsilon_t, \qquad \varepsilon_t = \exp(a x_t), \tag{27}$$

where ϵ_t is the measurement error, a is a real-valued scalar and $\{x_t\}$ is a stochastic process. The symbolism $\perp$ means $\{\sigma_t^2\}$ and $\{\epsilon_t^2\}$ are independent processes. This, together with $E(\epsilon_t) = 1$, implies that the volatility proxy is unbiased: $E(V_t) = E(\sigma_t^2)$ for all t. In the experiments, two classes of DGPs are studied:[5]

$$\begin{aligned}
\text{DGP 1:} \quad & a \in \{0, 0.2, 0.4\}, \qquad x_t \overset{iid}{\sim} N(0,1), \\
& \quad a = 0.0: \quad \phi_0 = 0.00, \quad \phi_1 = 1.00, \\
& \quad a = 0.2: \quad \phi_0 = 0.28, \quad \phi_1 = 0.72, \\
& \quad a = 0.4: \quad \phi_0 = 0.62, \quad \phi_1 = 0.38,
\end{aligned} \tag{28}$$

$$\begin{aligned}
\text{DGP 2:} \quad & a \in \{0.2, 0.4\}, \qquad x_t = 0.9 x_{t-1} + a e_t, \qquad e_t \overset{iid}{\sim} N(0,1), \\
& \quad a = 0.2: \quad \phi_0 = 0.07, \quad \phi_1 = 0.93, \\
& \quad a = 0.4: \quad \phi_0 = 0.58, \quad \phi_1 = 0.42.
\end{aligned} \tag{29}$$

In the first class of DGPs, ϵ_t is iid, and so $E(V_t|\mathcal{F}_{t-1}) = \sigma_t^2$ for all t. In the specific case where $a = 0$, there is no measurement error and so $\sigma_t^2 = V_t$ $a.s.$. When V_t measures σ_t^2 with error (i.e. $a > 0$), the null of the Standard MZ-test does not hold, since $\phi_0 \neq 0$ and $\phi_1 \neq 1$. In the second class of DGPs, ϵ_t is dependent and governed by a persistent AR(1) process in the exponent. Accordingly, while $E(V_t) = E(\sigma_t^2)$ by construction, conditional unbiasedness does not hold: $E(V_t|\mathcal{F}_{t-1}) \neq \sigma_t^2$.

The results of the simulations are contained in Table 3. When $a = 0$, then V_t measures σ_t^2 with no error. Both the Standard and Modified MZ-tests are notably oversized in this case, in particular in small samples where the discrepancy between the empirical and nominal sizes can be as large as 14%-points. For the Standard MZ-test, closer inspection of the simulation results reveals that the poor size is due to a finite sample bias in the estimates of ϕ_0 and ϕ_1. The Modified MZ-test is less affected by the finite sample bias, since the biases cancel each other out when computing their sum. Nevertheless, the best performance is exhibited by the Restricted MZ-test, since it is well-sized across the sample sizes studied. Indeed, already at $T = 500$ the discrepancy between the empirical and nominal size is less than 1%-point. Increasing the measurement error to $a = 0.2$ and $a = 0.4$ in DGP 1 confirms that the Standard MZ-test is flawed: As T increases, the probability of rejecting H_0 tends to 1. The size properties of the Modified and Restricted MZ-test, by contrast, improve as the sample size T increases. The improvement for the former is somewhat slow, since the discrepancy between the empirical and nominal sizes range from about 3 to 8 percentage points for $T = 1000$. For the Restricted MZ-test, by contrast, the discrepancy between the empirical and nominal size is again small and about 1%-point already when $T = 500$.

[5]The values of ϕ_0 and ϕ_1 when $a \neq 0$ are obtained by simulation.

The results of the DGP 2 simulations are similar: The Standard MZ-test is flawed in the presence of measurement error, the Modified and Restricted MZ-tests rectify the flaw, and the Restricted MZ-test has better empirical size across sample sizes when compared with the Modififed MZ-test. One notable difference compared with DGP 1, however, occurs when the measurement error becomes large, i.e. when $a = 0.4$. In this case, the Restricted MZ-test is generally undersized, and the discrepancy is increasing in T. A possible explanation is that increasing a in DGP 2 also strengthens the serial dependence of the measurement error ϵ_t. This may not be appropriately reflected in how the Newey and West (1987) coefficient-covariance is computed.

5 An illustration

To illustrate the ideas, tests and results of this paper, twelve volatility proxies used in three seminal studies are revisited. The three studies are: Andersen and Bollerslev (1998), Hansen and Lunde (2005), and Patton (2011). The data are freely available on the internet, and they all rely on a connection between their underlying notion of volatility and the expectation of squared return. Table 4 lists the volatility proxies and their samples. Note that the DM/USD proxy in Hansen and Lunde (2005) is the same as in Andersen and Bollerslev (1998) but divided by 0.8418, see Hansen and Lunde (2005, p. 881).

Table 5 contains the results of Tests 1–4 for identification, and an estimate of and test for bias (i.e. the Restricted MZ-test from Section 4). The p-values of Tests 1 and 2 suggest four out of twelve volatility proxies are not weakly identified at the 10% significance level: DM/USD1, IBM1, IBM65min and IBM5min. Their estimates of $\widehat{h}$ vary from 0.810 (DM/USD1) to 1.141 (IBM1). Tests 3 and 4 are implemented as tests for 1st. order autocorrelation in z_t^2 and $\ln z_t^2$, respectively. One or both p-values are less than 10% for five proxies: DM/USD1, DM/USD2, IBM65min, IBM15min and IBM5min. Interestingly, each of these five proxies exhibit a negative first order autocorrelation in z_t^2. While it is not always significant at 10%, it does suggest a log-MEM is more suitable as a model of h_t than a MEM of the GARCH-type, since the latter is not compatible with a negative first order autocorrelation in z_t^2. According to the Restricted MZ-test for bias, three of the proxies are biased for $E(\sigma_t^2)$ at the 10% level: DM/USD1, IBM65min and IBM5min.

As a minimum, a volatility proxy should satisfy weak identification if it is to be used as a substitute for an expectation of squared return. Table 6 contains the results of Tests 1–4 applied to the weakly corrected versions of DM/USD1, IBM1, IBM65min and IBM5min:

$$\begin{aligned} \text{DM/USD1:} &\quad \widehat{V}_t = \widehat{h}V_t, &\quad \widehat{h} = 0.810, \\ \text{IBM1:} &\quad \widehat{V}_t = \widehat{h}V_t, &\quad \widehat{h} = 1.141, \\ \text{IBM65min:} &\quad \widehat{V}_t = \widehat{h}V_t, &\quad \widehat{h} = 1.037, \\ \text{IBM5min:} &\quad \widehat{V}_t = \widehat{h}V_t, &\quad \widehat{h} = 0.902. \end{aligned}$$

Unsurprisingly, the corrected proxies satisfy weak identification at all significance levels. Interestingly, three of the four corrected proxies are also less biased. The exception is IBM1, whose bias is larger after the correction.

A total of five proxies do not satisfy semi-strong identification. To correct them for semi-strong identification, a log-MEM(1,0) specification of h_t is fitted to z_t^2 for each of them. The reasons a log-MEM(1,0) is chosen are two. First, according to Proposition 1 there exists a log-MEM(1,0) representation under general and mild assumptions. Second, the log-MEM(1,0) provides a better fit than a log-MEM(1,1) according to both the Schwarz (1978) and Akaike (1974) information criteria. This leads to the following five corrected proxies:

$$\text{DM/USD1:} \qquad \widehat{V}_t = \widehat{h}_t V_t, \qquad \ln \widehat{h}_t = 0.3508 - \underset{(0.0618)}{0.1030} \ln z_{t-1}^2 \tag{30}$$

$$\text{DM/USD2:} \qquad \widehat{V}_t = \widehat{h}_t V_t, \qquad \ln \widehat{h}_t = -0.1609 - \underset{(0.0618)}{0.1030} \ln z_{t-1}^2 \tag{31}$$

$$\text{IBM65min:} \qquad \widehat{V}_t = \widehat{h}_t V_t, \qquad \ln \widehat{h}_t = 0.7498 + \underset{(0.0190)}{0.0597} \ln z_{t-1}^2 \tag{32}$$

$$\text{IBM15min:} \qquad \widehat{V}_t = \widehat{h}_t V_t, \qquad \ln \widehat{h}_t = 0.7220 + \underset{(0.0190)}{0.0613} \ln z_{t-1}^2 \tag{33}$$

$$\text{IBM5min:} \qquad \widehat{V}_t = \widehat{h}_t V_t, \qquad \ln \widehat{h}_t = 0.7156 + \underset{(0.0190)}{0.0667} \ln z_{t-1}^2 \tag{34}$$

Next, Tests 1 – 4 are applied to $\widehat{z}_t^2 = r_t^2/\widehat{V}_t$, together with the Restricted MZ-test for bias. Table 7 contains the results. The corrected proxies satisfy both weak and semi-strong identification at the 10% significance level, since all the p-values are larger than 0.22. Interestingly, however, the bias is not always reduced. Indeed, only for DM/USD1 is it reduced, and for IBM65min, IBM15min and IBM5min it increases notably. This provides an example of the trade-off between the kind of identification that is sought, and the extent of the resulting bias.

A total of six proxies did not satisfy either weak or semi-strong identification, or both. All-in-all, we may conclude that four of these (DM/USD1, DM/USD2, IBM65min and IBM5min) should be corrected, the conclusion is not clear-cut for one proxy (IBM1), and one proxy should not be corrected (IBM15min). DM/USD1 should be corrected to satisfy semi-strong identification, since this provides the best improvement according to both identification and bias. Correcting the DM/USD2 proxy so that it satisfies semi-strong identification improves $\widehat{h}$ from 0.962 to 1.000, but worsens the bias from 0.000 to 0.015. However, the deterioration in bias is marginal, and the resulting bias is insignificantly different from zero at common significance levels. So the overall conclusion is that it should be corrected for semi-strong identification. The results suggest IBM65min and IBM5min should be corrected to satisfy weak identification, since this also reduces the bias. They should not be corrected to satisfy semi-strong identification, since this induces a substantial bias. It is not clear-cut that the IBM1 proxy should be corrected to satisfy weak identification. While the correction improves $\widehat{h}$ substantially from 1.141 to 1.000, the bias is worsened notably from 0.000 to -0.844. Finally, the IBM15min proxy, which is already weakly identified, should not be corrected for semi-strong identification, since this induces a notable bias.

6 Conclusions

A tripartite distinction between strong, semi-strong and weak identification of a volatility proxy as an expectation of squared returns is introduced. Strong identification implies semi-strong identification, and semi-strong identification implies weak identification. However, their converses are not true. The notions of identification and unbiasedness differ. The former is multiplicative, whereas the latter is additive. This means a biased proxy can be identified, and an unbiased proxy can fail to be identified. For meaningful use of a volatility proxy as a substitute for an expectation of squared return in volatility forecast evaluation, the proxy should – as a minimum – be weakly identified as an expectation. Otherwise, the proxy is not on a comparable scale. The multiplicative transformation at the base of the definition implies that well-known tests and procedures can be used to check and correct for identification. Monte Carlo simulations verify that the tests are well-sized and powerful in finite samples. Specifications of h_t for identification correction is discussed. It is shown that, subject to mild and general assumptions, there exists a log-MEM(p,0) representation that can be estimated by a least squared procedure. This means a general but flexible and straightforward procedure for correction is, in general, available. Next, it is shown that the Standard MZ-test is, in general, flawed when the proxy measures σ_t^2 with error. Straightforward modifications that rectifies the flaw are derived, and Monte Carlo simulations show that the simplest of them is particularly well-sized. Finally, in an empirical illustration, twelve volatility proxies from three seminal studies are revisited. Half of them are found to not satisfy either semi-strong or weak identification, but their corrected counterparts do. However, identification correction does not always lead to a reduction in bias, thus illustrating the tradeoff between the chosen specification of h_t and the resulting bias.

References

Aït-Sahalia, Y. and P. Mykland (2009). Estimating Volatility in the Presence of Market Microstructure Noise: A Review of the Theory and Practical Considerations. In T. Andersen, R. Davis, J.-P. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*. Berlin: Springer.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control 19*, 716–723.

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

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

Bandi, F. M. and J. Russell (2008). Market microstructure noise, realized variance and optimal sampling. *Review of Economic Studies 75*, 339–369.

Bollerslev, T., A. Patton, and R. Quaedvlieg (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics 192*, 1–18.

Brownlees, C., F. Cipollini, and G. Gallo (2012). Multiplicative Error Models. In L. Bauwens, C. Hafner, and S. Laurent (Eds.), *Handbook of Volatility Models and Their Applications*, pp. 223–247. New Jersey: Wiley.

Drost, F. C. and T. E. Nijman (1993). Temporal Aggregation of Garch Processes. *Econometrica 61*, 909–927.

Duan, N. (1983). Smearing Estimate: A Nonparametric Retransformation Method. *Journal of the Americal Statistical Association 78*, pp. 605–610.

Engle, R. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflations. *Econometrica 50*, 987–1008.

Engle, R. F. and J. R. Russell (1998). Autoregressive Conditional Duration: A New Model of Irregularly Spaced Transaction Data. *Econometrica 66*, 1127–1162.

Francq, C. and J.-M. Zakoïan (2019). *GARCH Models*. New York: Wiley. 2nd. Edition.

Hansen, P. R. and A. Lunde (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics 20*, 873–889.

Hautsch, N., P. Malec, and M. Schienle (2013). Capturing the zero: a new class of zero-augmented distributions and multiplicative error processes. *Journal of Financial Econometrics 12*, 89–121.

Ljung, G. and G. Box (1979). On a Measure of Lack of Fit in Time Series Models. *Biometrika 66*, 265–270.

Mincer, J. and V. Zarnowitz (1969). The Evaluation of Economic Forecasts. In J. Zarnowitz (Ed.), *Economic Forecasts and Expectations*, pp. 3–46. New York: National Bureau of Economic Research.

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

Park, S. and O. Linton (2012). Realized Volatility: Theory and Applications. In L. Bauwens, C. Hafner, and S. Laurent (Eds.), *Handbook of Volatility Models and Their Applications*, pp. 319–345. New Jersey: Wiley.

Patton, A. J. (2011). Volatility Forecast Evaluation and Comparison Using Imperfect Volatility Proxies. *Journal of Econometrics 160*, 246–256. Code and data: http://econ.duke.edu/~ap172/Patton_robust_loss_apr06.zip.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*, 461–464.

Sucarrat, G. (2009). Forecast Evaluation of Explanatory Models of Financial Variability. *Economics – The Open-Access, Open-Assessment E-Journal 3*. http://www.economics-ejournal.org/economics/journalarticles/2009-8.

Sucarrat, G. (2019). The Log-GARCH Model via ARMA Representations. In J. Chevallier, S. Goutte, D. Guerreiro, S. Saglio and B. Sanhadji (eds.): *Financial Mathematics, Volatility and Covariance Modelling, Volume 2*. Working Paper version: https://mpra.ub.uni-muenchen.de/id/eprint/100386.

Sucarrat, G., S. Grønneberg, and Á. Escribano (2016). Estimation and Inference in Univariate and Multivariate Log-GARCH-X Models When the Conditional Density is Unknown. *Computational Statistics and Data Analysis 100*, 582–594.

Violante, F. and S. Laurent (2012). Volatility Forecasts Evaluation and Comparison. In L. Bauwens, C. Hafner, and S. Laurent (Eds.), *Handbook of Volatility Models and Their Applications*, pp. 465–486. New Jersey: Wiley.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity. *Econometrica 48*, 817–838.

Yeh, J.-H. and J.-N. Wang (2019). Bias-corrected realized variance. *Econometric Reviews 38*, 170–192.

Table 1: Rejection frequencies (in %) of Tests 1 and 2 in Section 2.3

ID	DGP	T	Test 1: 10%	Test 1: 5%	Test 1: 1%	Test 2: 10%	Test 2: 5%	Test 2: 1%
1	$h_t = 1.00$:	250	11.29	6.18	1.68	10.76	5.67	1.36
		500	10.75	5.62	1.51	10.50	5.52	1.26
		1000	10.31	5.25	1.29	10.48	5.28	1.28
		2000	9.67	4.55	1.14	10.56	5.28	1.19
		5000	10.18	4.90	1.12	10.00	5.17	1.06
	$h_t = 0.90$:	250	38.39	28.54	15.00	33.57	23.44	9.68
		500	55.92	44.68	25.19	52.29	40.02	19.49
		1000	78.08	68.37	47.87	76.56	65.92	42.27
		2000	96.00	92.64	80.71	95.52	91.75	78.61
		5000	99.99	99.98	99.68	99.98	99.96	99.61
	$h_t = 1.10$:	250	25.27	14.88	3.85	29.15	19.06	7.01
		500	40.55	28.41	10.18	44.19	31.93	14.04
		1000	66.29	53.62	27.33	68.34	56.75	33.00
		2000	90.25	83.31	62.44	91.12	85.45	67.03
		5000	99.87	99.73	98.32	99.87	99.74	98.16
2a)	$\boldsymbol{\theta} = (-0.16, -0.1)'$:	250	10.95	5.86	1.92	11.01	5.23	1.16
		500	10.73	5.79	1.44	10.64	5.40	1.14
		1000	10.42	5.32	1.32	10.23	4.95	1.09
		2000	9.90	4.91	1.14	9.32	4.60	0.98
		5000	10.15	5.15	1.04	9.32	4.56	0.82
2b)	$\boldsymbol{\theta} = (0, -0.1)'$:	250	44.49	30.41	9.86	49.02	36.69	16.72
		500	71.42	58.69	31.18	75.47	64.71	39.50
		1000	94.54	89.60	71.92	95.22	90.87	76.44
		2000	99.85	99.56	97.58	99.94	99.76	98.29
		5000	100.00	100.00	100.00	100.00	100.00	100.00
2c)	$\boldsymbol{\theta} = (0, 0.1)'$:	250	40.81	31.71	17.76	36.00	25.93	11.86
		500	56.43	45.41	27.46	53.53	41.53	20.78
		1000	80.20	71.26	49.84	77.76	67.29	43.67
		2000	96.08	92.91	81.94	95.42	91.91	78.01
		5000	99.98	99.94	99.74	99.97	99.96	99.54

Rejection frequencies for significance levels 10%, 5% and 1%. 10 000 simulations.

Table 2: Rejection frequencies (in %) of Tests 3 and 4 in Section 2.3

			Test 3:			Test 4:		
ID	DGP	T	10%	5%	1%	10%	5%	1%
1	$h_t = 1.00$:	250	9.27	4.55	0.99	9.82	4.78	1.11
		500	9.56	4.99	1.09	9.83	4.81	0.90
		1000	9.78	4.85	1.13	10.45	5.10	1.16
		2000	9.52	4.78	0.96	10.03	5.12	1.11
		5000	9.94	4.73	0.95	9.64	4.86	0.92
2a)	$\boldsymbol{\theta} = (-0.16, -0.1)'$	250	44.85	28.17	6.28	50.13	37.83	17.62
		500	73.73	58.54	26.44	73.96	63.11	38.98
		1000	95.06	89.63	69.39	93.48	88.38	72.76
		2000	99.91	99.63	97.51	99.80	99.38	97.23
		5000	100.00	100.00	100.00	100.00	100.00	100.00
2b)	$\boldsymbol{\theta} = (0, -0.1)'$	250	44.36	27.60	6.52	51.17	37.90	17.02
		500	73.42	58.55	26.54	73.89	63.58	39.51
		1000	95.17	90.18	69.60	93.60	88.93	73.34
		2000	99.85	99.60	97.49	99.68	99.27	97.31
		5000	100.00	100.00	99.99	100.00	100.00	100.00
2c)	$\boldsymbol{\theta} = (0, 0.1)'$	250	40.26	30.45	15.46	43.00	31.40	13.82
		500	63.52	52.51	31.85	70.85	58.83	34.55
		1000	88.39	81.32	63.03	93.36	87.86	70.30
		2000	99.34	98.55	93.75	99.76	99.38	96.84
		5000	100.00	100.00	99.99	100.00	100.00	100.00

Rejection frequencies for significance levels 10%, 5% and 1%. 10 000 simulations.

Table 3: Rejection frequencies (in %) of the tests for bias in Section 4

DGP		T	Standard MZ-test 10%	5%	1%	Modified MZ-test 10%	5%	1%	Restricted MZ-test 10%	5%	1%
1	$a = 0.0$:	250	23.89	16.94	8.30	17.14	11.07	5.09	11.95	7.03	2.55
		500	20.49	13.93	6.40	14.23	8.71	3.10	10.87	5.88	1.84
		1000	17.49	11.23	4.52	12.57	7.20	2.10	9.92	5.51	1.50
		2000	16.18	10.03	4.08	11.73	6.26	1.59	9.95	5.07	1.25
		5000	14.01	8.13	2.77	10.86	5.44	1.40	9.84	5.05	0.97
	$a = 0.2$:	250	53.28	43.94	28.95	22.33	15.31	7.36	12.30	7.29	2.59
		500	56.72	47.67	32.45	20.17	13.22	5.51	11.01	6.11	1.87
		1000	65.06	56.27	40.75	17.96	11.40	4.28	10.83	5.39	1.47
		2000	75.01	67.60	52.80	17.14	10.41	3.68	10.62	5.71	1.25
		5000	87.70	83.25	73.25	16.04	9.54	2.96	10.37	5.39	1.27
	$a = 0.4$:	250	90.46	86.16	76.55	28.22	20.48	10.34	11.85	6.44	2.27
		500	94.90	92.93	87.09	25.74	17.90	8.31	11.11	5.96	1.72
		1000	98.12	97.03	93.94	23.67	16.22	7.21	10.44	5.45	1.58
		2000	99.24	98.82	97.83	22.24	14.53	5.99	10.50	5.28	1.34
		5000	99.81	99.66	99.40	21.53	14.18	5.56	9.94	5.21	1.30
2	$a = 0.2$:	250	32.66	24.24	13.17	19.78	13.41	6.23	11.74	6.74	2.28
		500	30.29	22.65	11.33	16.19	10.37	3.96	11.47	6.45	2.14
		1000	29.11	21.28	10.55	14.31	8.85	3.02	10.76	5.84	1.68
		2000	29.10	21.36	10.72	13.01	7.19	2.11	10.27	5.26	1.36
		5000	32.62	24.12	12.71	11.81	6.11	1.70	10.25	5.01	1.01
	$a = 0.4$:	250	86.88	82.31	71.87	28.38	20.71	11.06	8.20	3.97	1.08
		500	93.18	90.45	82.66	25.35	17.41	8.42	7.22	3.59	0.76
		1000	96.53	95.20	91.28	23.07	15.62	6.74	6.29	2.92	0.48
		2000	98.74	98.17	96.51	21.43	14.22	5.55	5.76	2.37	0.29
		5000	99.56	99.34	98.80	20.04	13.00	4.74	5.63	2.37	0.34

Rejection frequencies for significance levels 10%, 5% and 1%. 10 000 replications.

Table 4: List of studies (see Section 5)

Study	Proxy(/ies)	Period	T
Andersen and Bollerslev (1998):	DM/USD	1/10/1992 – 29/9/1993	260
Hansen and Lunde (2005):	DM/USD	1/10/1992 – 29/9/1993	260
	IBM	1/6/1999 – 31/5/2000	254
Patton (2011):	IBM	4/1/1993 – 31/12/2003	2772

Table 5: Identification tests of volatility proxies (see Section 5)

	Proxy	Test 1 $\widehat{h}$ [$p-val$]	Test 2 $\ln\widehat{h}$ [$p-val$]	Test 3 $\widehat{\rho}_1(z_t^2)$ [$p-val$]	Test 4 $\widehat{\rho}_1(\ln z_t^2)$ [$p-val$]	Bias [$p-val$]
Andersen and Bollerslev (1998):	DM/USD1	0.810 [0.000]	−0.211 [0.001]	−0.151 [0.014]	−0.103 [0.095]	−0.101 [0.010]
Hansen and Lunde (2005):	DM/USD2	0.962 [0.523]	−0.038 [0.531]	−0.151 [0.014]	−0.103 [0.095]	0.000 [1.000]
	IBM1	1.141 [0.099]	0.132 [0.078]	0.016 [0.803]	0.014 [0.817]	0.000 [1.000]
	IBM2	1.047 [0.557]	0.046 [0.548]	0.045 [0.476]	0.011 [0.864]	0.000 [1.000]
	IBM3	1.041 [0.603]	0.040 [0.595]	0.033 [0.600]	0.009 [0.882]	0.000 [1.000]
	IBM4	1.082 [0.329]	0.079 [0.310]	0.030 [0.630]	0.017 [0.790]	0.000 [1.000]
	IBM5	1.083 [0.324]	0.080 [0.305]	0.022 [0.722]	0.016 [0.792]	0.000 [1.000]
	IBM6	1.008 [0.916]	0.008 [0.915]	0.026 [0.678]	0.015 [0.813]	0.000 [1.000]
	IBM7	1.006 [0.938]	0.006 [0.938]	0.021 [0.738]	0.012 [0.847]	0.000 [1.000]
Patton (2011):	IBM 65min	1.037 [0.049]	0.036 [0.045]	−0.042 [0.027]	0.060 [0.002]	0.314 [0.005]
	IBM 15min	1.017 [0.456]	0.017 [0.453]	−0.026 [0.179]	0.061 [0.001]	0.118 [0.381]
	IBM 5min	0.902 [0.000]	−0.103 [0.000]	−0.029 [0.123]	0.067 [0.000]	−0.291 [0.028]

$\widehat{h}$, sample average of z_t^2. $\ln\widehat{h}$, natural log of $\widehat{h}$. $\widehat{\rho}_1(z_t^2)$, first order sample autocorrelation of z_t^2. $\widehat{\rho}_1(\ln z_t^2)$, first order sample autocorrelation of $\ln z_t^2$. $p-val$, p-value of test. Test 1, $H_0 : h = 1$ vs. $H_A : h \neq 1$, see (11). Test 2, $H_0 : \ln h = 0$ vs. $H_A : \ln h \neq 0$, see (12). Test 3, Ljung and Box (1979) test for first order autocorrelation in z_t^2, see (13). Test 4, Ljung and Box (1979) test for first order autocorrelation in $\ln z_t^2$, see (14). Bias, Restricted MZ-test, see (26), where the estimated bias is computed as $T^{-1}\sum_{t=1}^{T}(r_t^2 - V_t)$.

Table 6: Weak identification of volatility proxies (see Section 5)

	Proxy	Test 1 $\widehat{h}$ [$p-val$]	Test 2 $\ln\widehat{h}$ [$p-val$]	Test 3 $\widehat{\rho}_1(\widehat{z}_t^2)$ [$p-val$]	Test 4 $\widehat{\rho}_1(\ln\widehat{z}_t^2)$ [$p-val$]	Bias [$p-val$]
Andersen and Bollerslev (1998):	DM/USD1	1.000 [1.000]	0.000 [1.000]	−0.151 [0.014]	−0.103 [0.095]	0.020 [0.596]
Hansen and Lunde (2005):	IBM1	1.000 [1.000]	0.000 [1.000]	0.016 [0.803]	0.014 [0.817]	−0.844 [0.109]
Patton (2011):	IBM 65min	1.000 [1.000]	0.000 [1.000]	−0.042 [0.027]	0.060 [0.002]	0.156 [0.162]
	IBM 5min	1.000 [1.000]	0.000 [1.000]	−0.029 [0.123]	0.067 [0.000]	0.189 [0.156]

Tests 1–4 are of $\widehat{z}_t^2 = r_t^2/\widehat{V}_t$, where $\widehat{V}_t = \widehat{h}_t V_t$ is the identification corrected proxy. $\widehat{h}$, sample average of $\widehat{z}_t^2$. $\widehat{\rho}_1(\widehat{z}_t^2)$, first order sample autocorrelation of $\widehat{z}_t^2$. $\widehat{\rho}_1(\ln\widehat{z}_t^2)$, first order sample autocorrelation of $\ln\widehat{z}_t^2$. $p-val$, p-value of test. Test 1, $H_0 : h = 1$ vs. $H_A : h \neq 1$, see (11). Test 2, $H_0 : \ln h = 0$ vs. $H_A : \ln h \neq 0$, see (12). Test 3, Ljung and Box (1979) test for first order autocorrelation in z_t^2, see (13). Test 4, Ljung and Box (1979) test for first order autocorrelation in $\ln z_t^2$, see (14). Bias, Restricted MZ-test, see (26), where the estimated bias is computed as $T^{-1}\sum_{t=1}^{T}(r_t^2 - \widehat{V}_t)$.

Table 7: Semi-strong identification of volatility proxies (see Section 5)

	Proxy	Test 1 $\widehat{h}$ [$p-val$]	Test 2 $\ln\widehat{h}$ [$p-val$]	Test 3 $\widehat{\rho}_1(\widehat{z}_t^2)$ [$p-val$]	Test 4 $\widehat{\rho}_1(\ln\widehat{z}_t^2)$ [$p-val$]	Bias [$p-val$]
Andersen and Bollerslev (1998):	DM/USD1	1.000 [1.000]	0.000 [1.000]	−0.075 [0.222]	−0.014 [0.818]	0.015 [0.719]
Hansen and Lunde (2005):	DM/USD2	1.000 [1.000]	0.000 [1.000]	−0.075 [0.222]	−0.014 [0.818]	0.015 [0.719]
Patton (2011):	IBM 65min	1.000 [1.000]	0.000 [1.000]	−0.023 [0.228]	−0.001 [0.948]	−4.086 [0.000]
	IBM 15min	1.000 [1.000]	0.000 [1.000]	−0.020 [0.281]	−0.002 [0.933]	−4.142 [0.000]
	IBM 5min	1.000 [1.000]	0.000 [1.000]	−0.017 [0.375]	−0.002 [0.919]	−4.737 [0.000]

The tests are of $\widehat{z}_t^2 = r_t^2/\widehat{V}_t$, where $\widehat{V}_t = \widehat{h}_t V_t$ is the identification corrected proxy. $\widehat{h}$, sample average of $\widehat{z}_t^2$. $\widehat{\rho}_1(\widehat{z}_t^2)$, first order sample autocorrelation of $\widehat{z}_t^2$. $\widehat{\rho}_1(\ln\widehat{z}_t^2)$, first order sample autocorrelation of $\ln\widehat{z}_t^2$. $p-val$, p-value of test. Test 1, $H_0 : h = 1$ vs. $H_A : h \neq 1$, see (11). Test 2, $H_0 : \ln h = 0$ vs. $H_A : \ln h \neq 0$, see (12). Test 3, Ljung and Box (1979) test for first order autocorrelation in z_t^2, see (13). Test 4, Ljung and Box (1979) test for first order autocorrelation in $\ln z_t^2$, see (14).