On Heteroskedasticity and Regimes in Volatility Forecasting

Fabrizio Cipollini
DiSIA, Università di Firenze, Italy
e-mail: cipollini@disia.unifi.it

Giampiero M. Gallo*
DiSIA, Università di Firenze, Italy
e-mail: gallog@disia.unifi.it

Edoardo Otranto
Università di Messina, Italy
e-mail: eotranto@unime.it

This Version: September 14, 2017

Abstract

In this paper we discuss some deep implications of the recent paper by Bollerslev *et al.* (2016) (BPQ). In BPQ the volatility dynamics modeled as a HAR is augmented by a term involving quarticity in order to correct measurement errors in realized variance. We show that the model is observationally equivalent to another where a quadratic term in realized variance accounts for a faster mean reversion when volatility is high. We argue that heteroskedasticity (volatility of volatility) and a time-varying mean seem to play a role of higher order of importance than measurement errors. In fact, the quarticity/quadratic terms disappear within an AMEM, and more so when a Markov Switching dynamics is considered. Some simulation results show that when the DGP is a (MS-)AMEM, such terms turn out spuriously significant in a HAR. Forecast performance of the (MS-)AMEM is superior to the augmented HARs.

Keywords: Realized volatility, Forecasting, Measurement errors, HAR, AMEM, Markov switching, Volatility of volatility
JEL codes: C22, C51, C53, C58

*Corresponding author: Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti" (DiSIA), Università di Firenze, Viale G.B. Morgagni, 59 - 50134 Firenze – Italy; ph. +39 055 2751 591. The authors thank Fulvio Corsi and participants in the 2016 CFE Conference in Seville.

1 Introduction

Financial market volatility exhibits persistence, be it implicit in option prices (as in VIX Whaley, 2009), estimated through ultra–high frequency data (UHFD – any of the several flavors of realized variance and daily range, cf. Andersen *et al.*, 2006; Alizadeh *et al.*, 2002), or derived as the conditional (one–step ahead) variance of returns in a GARCH–type model (cf. Teräsvirta, 2009). While in the GARCH modeling, measurement and prediction coincide, when volatility measurement is UHFD–based, a separate modeling strategy has to be set up, invariably grounded on some sort of an autoregressive scheme. The choice is by now abundant: in general, models aim at capturing empirical regularities in the series, mainly persistence, a slowly moving *average* level of volatility, jumps, the presence of regimes, and possible long memory features.

The fact that realized volatility, whatever the method, is not *volatility*, but an estimate of it, has raised the point that, in spite of it being consistent, the measurement error in the estimator relative to the *true* value may spoil whatever modeling procedure is adopted. This point is forcefully pursued by Bollerslev *et al.* (2016) (henceforth, BPQ) who illustrate the theoretical consequences of modeling and forecasting realized variance. As a textbook case of errors–in–variables, they make reference to a simple AR(1) model to show that using realized variance *in lieu* of its theoretical counterpart leads to a downward bias in the estimated autoregressive coefficient; such a bias is a function of the integrated quarticity. In other words, in practice, what we measure leads to an estimated forecasting model which is more persistent than it should be. The framework chosen by BPQ to illustrate a possible correction for this measurement error is the Heterogeneous Autoregressive model (HAR, Corsi, 2009) a linear model with additive errors, where dependence on the past, aggregated at lower frequencies, was found to mimic long memory features. Within this HAR, BPQ correct for the measurement error effect by inserting an interaction term between the lagged realized variance and the (square root of) realized quarticity. A negative sign of the related coefficient ensures that the higher the measurement error the lower the persistence of the past into the presence.

Theoretical results on measurement errors aside, in this paper we address the issue from the viewpoint that accounting for heteroskedasticity and for the presence of slow moving underlying low frequency trend in the data makes the importance of measurement errors of a lesser order. We start from the stylized fact that (sqrt-)quarticity is strongly correlated with realized variance (in our panel of 29 assets, the median is 0.934). Therefore, if we replace the interaction with a curvature term (realized variance squared), we get similar results (negative coefficient, mostly significant). We have therefore a competing – observationally equivalent – model where the relevant nonlinear term added to the base HAR model reduces persistence when realized variance is high. The explanation has economic content: high levels of lagged realized variance imply a faster absorption of news and a faster reversion to the mean. The fundamental questions addressed in this paper are

1. whether the base HAR is a well specified model, or by inserting interaction or quadratic terms we are rather detecting heteroskedasticity *à la* White (1980). This is a point which confirms the findings of Corsi *et al.* (2008) about the need to account for volatility of volatility (i.e. heteroskedasticity) within a HAR-GARCH. As an alter-

native to HAR or HAR-GARCH, we can consider Multiplicative Error Models (MEMs Engle, 2002), which are heteroskedastic in nature;

2. whether the level which the series oscillates around is a constant term (the unconditional mean), or, following the seminal paper by Engle and Rangel (2008), it is a low frequency component. Neglecting this piece of dynamics may introduce a further element of misspecification, no matter whether it is modeled as a flexible function of time, as in Engle and Rangel (2008), or regime specific dynamics, as in MS-GARCH(Dueker, 1997; Haas *et al.*, 2004), or a smooth transition function dependent on a state variable, as in Amado and Teräsvirta (2014) in GARCH models.

In modeling realized variance or volatility directly, and citing the presence of measurement errors in realized variance, Maheu and McCurdy (2002) use a Markov Switching ARMAX model where regime dependency is found both in the conditional mean and the conditional variance of the realized volatility. Parallel paths have been followed, by Brownlees and Gallo (2010) or Brownlees and Gallo (2011) for splines, Gallo and Otranto (2015) for Markov Switching (MS-AMEM) and Smooth Transition (ST-AMEM) MEMs. Further refinements are worth mentioning, such as in HAR the consideration of jumps as in Andersen *et al.* (2007), or of a smooth transition as in McAleer and Medeiros (2008) or the possibility of combining HAR and MEMs with Markov Switching as in Gallo and Otranto (2015).

When we estimate (univariate) asymmetric MEMs (AMEM) on the same panel of data, we find that the interaction and curvature terms are much less significant, and that having the short run dynamics about a regime–specific mean in a MS-AMEM kills the relevance of such terms even more. Interestingly, when we simulate data from an AMEM and an MS-AMEM process (without interaction or curvature terms) and then we estimate the HAR model with such terms, we find a strong evidence of them being individually statistically significant. The conclusion we suggest is that the evidence of curvature within the HAR class of models found by BPQ not only could be attributed to alternative explanations (higher variances induce a faster mean reversion, rather than measurement errors) but it may be a sign of misspecification. Following the evidence that HAR is misspecified, a sounder multiplicative specification (AMEM) accounts for heteroskedasticity and does not find a strong evidence for the extra terms. The further (MS-AMEM) refinement with short term dynamics around a regime specific mean eliminates the evidence of a curvature. These results are strenghtened by the analysis of prediction results both in– and out–of–sample which show that, while HAR-GARCH is a competitive model, AMEM, but more so, MS-AMEM provide better realized variance forecasts than HAR. This better out–of–sample performance is not a trivial result, also in view of the Hansen (2010) discussion about the trade-off between prediction ability and complexity of the model.

The structure of the paper is as follows: in Section 2 we recall the main theoretical points of realized variance being affected by measurement errors relative to integrated variance and the approach by BPQ to correct for them. In Section 3 we discuss the unavoidable observational equivalence between the augmented HAR by BPQ with an interaction term involving (sqrt-) quarticity and another HAR with a curvature term (the square of realized variance) suggesting a different rate of mean reversion in the presence of high volatility. In Section 4 we discuss a wider range of issues involving heteroskedasticity, asymmetric behavior in the presence of negative past returns, and regimes: we present three classes of models, the HAR-GARCH, the AMEM, and the MS-AMEM, which variously address these issues. A simulation study shows that when data are generated with AMEM or MS-AMEM (without extra terms), estimation by HAR generally gives significant interaction or curvature terms, while for the HAR-GARCH it does not. Section 6 provides evidence both in– (6.1) and out–of–sample (6.2) across all models. Concluding remarks follow.

2 Measurement Errors in Realized Variance

The BPQ paper considers a standard framework for the evolution of prices as a diffusion process in continuous time, recalling that the Integrated Variance (IV_t) is not directly observable but can be measured by one of the several versions of realized variance (RV_t) using ultra–high frequency data (Andersen and Benzoni, 2009). This implies the measurement error relationship (cf. Barndorff-Nielsen and Shephard (2002) for definitions and details)

$$RV_t = IV_t + \eta_t$$

where, conditional on the Integrated Quarticity IQ_t, $(\eta_t|IQ_t)$ has a normal distribution with mean zero and variance equal to $2\Delta IQ_t$. IQ_t, in turn, can be consistently estimated by a realized measure called *Realized Quarticity* (RQ_t, related to the fourth power of intradaily returns). In the BPQ discussion, forecasting IV_t with RV_t available builds both on the persistent behavior of IV_t and on the serial uncorrelatedness of η_t bringing forth that an AR(1) for IV_t would correspond to an ARMA(1,1) for RV_t. By contrast, if one estimated an AR(1) on RV_t, the autoregressive coefficient would be affected by a measurement error attenuation bias (cf. Equation (9) in the original BPQ paper, which we refer to for assumptions and details). For all practical purposes, an important consequence is that, even in this illustrative AR(1) framework for RV_t, the autoregressive coefficient is time–varying and related to the integrated quarticity, with the qualitative feature that, when the latter is large, the realized variance predictability loses its grip (the coefficient goes to zero).

In view of this result, BPQ suggest that the simplified AR(1) specification for RV_t be modified to include a correction term[1]

$$RV_t = \omega + (\alpha_1 + \alpha_{1Q} RQ_{t-1}^{1/2}) RV_{t-1} + \varepsilon_t = \omega + \alpha_{1,t-1} RV_{t-1} + \varepsilon_t,$$

where ε is a zero mean i.i.d. innovation term with assumed constant variance. With an assumed $\alpha_{1Q} < 0$, such an expression reproduces the feature mentioned above of giving less importance to days where RQ_{t-1} is high: in fact, such days would be the ones in which uncertainty in RV due to measurement error is the highest.

In order to make the analysis more realistic, BPQ adopt the HAR modeling framework (Corsi, 2009) for RV_t, enlarged to an HARQ (Equation (13) in BPQ) by including a component (dependent on the square root of the quarticity – hence the suffix Q) aimed at

[1] We prefer to index the time–varying coefficient by $t-1$ to highlight its $t-1$–measurability.

capturing the measurement error:

$$RV_t = \omega + (\underbrace{\alpha_D + \alpha_E RQ_{t-1}^{1/2}}_{\alpha_{1,t-1}})RV_{t-1} + \alpha_W \overline{RV}_{t-(2:5)} + \alpha_M \overline{RV}_{t-(6:22)} + \varepsilon_t \qquad (1)$$

where $\overline{RV}_{t-(h:k)}$ is the mean of $t-h$ to $t-k$ lagged RV values, and

$$RQ_t = \frac{M}{3}\sum_{i=1}^{M} r_{t,i}^4$$

is the realized quarticity ($r_{t,i}$ = return of the i-th intradaily bin, $i = 1, \ldots, M$). α_E is assumed non-positive so that $\alpha_{1,t} \leq \alpha_D$ and non–increasing in $RQ_{t-1}^{1/2}$. Note that in our specification, for better interpretation purposes, we prefer to avoid overlaps among regressors (hence lag 1 is not included in the weekly average, and lags 1 to 5 are not included in the monthly one). Moreover, for future reference, estimation is carried out by OLS and, as such, it delivers better fitting properties when a squared in–sample loss is concerned.

In their empirical results obtained on data for the S&P500 and 27 constituents of the Dow Jones Index, BPQ find evidence of negative significant α_E, and of an improved forecasting performance of the HARQ over HAR.

3 Two Observationally Equivalent Augmented HAR's

We consider a similar dataset as BPQ: we examine 28 components of the DJ30 Industrial Index plus the SPY ETF [2] over the period Jan. 3, 2003 to Dec. 31, 2015 (3273 observations total). We build the series of realized variances from TAQ data, cleaned according to the Brownlees and Gallo (2006) procedure. The first 2015 observations (up to Dec. 31, 2010) serve for the in–sample analysis, while we use the rest for an out–of–sample rolling analysis: we estimate on a 8–year period (starting from 2003-2010) and we produce one–step ahead predictions for the following year. As a leading example we will work on the SPY ETF which mimics the behavior of the Standard and Poor's 500.

Let us rewrite the HARQ specification (1) as

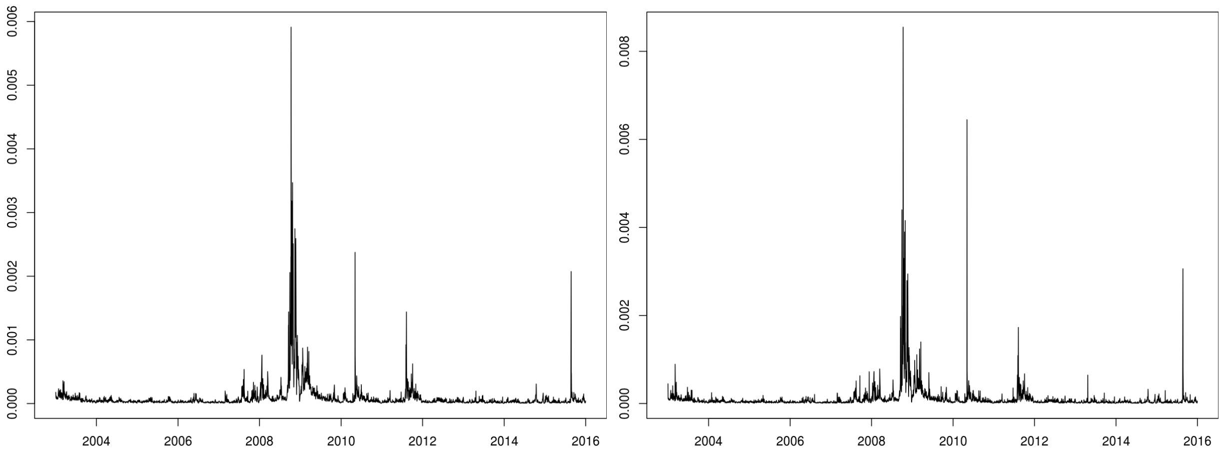
$$RV_t = \omega + \alpha_D RV_{t-1} + \alpha_E (RQ_{t-1}^{1/2} RV_{t-1}) + \alpha_W \overline{RV}_{t-(2:5)} + \alpha_M \overline{RV}_{t-(6:22)} + \varepsilon_t. \qquad (2)$$

The heart of this model lies on the extra term $RQ_{t-1}^{1/2} RV_{t-1}$ with expected coefficient $\alpha_E < 0$. This means that the reaction to the lagged realized variance is smaller (less persistence) in the presence of an interaction term between (the square root of) the realized quarticity and the realized variance, which should capture the presence of measurement errors in the Bollerslev *et al.* (2016) approach. This being an interaction term, the question rises as to

[2] Cf. Table 12 in the Appendix for the list of tickers of realized variances available over the entire sample period. Two components currently present in the index were excluded, since Travelers entered the DJ30 in 2009 and Visa in 2013.

whether the square root of the realized quarticity has substantial *additional* information relative to the realized variance, or whether it is high when the realized variance is high and vice versa. The similarity of the two series can be appreciated graphically in reference to the SPY ETF, as in Figure 1, where the realized variance is shown next to the corresponding square root of the quarticity.

Figure 1: SPY over the period Jan. 3, 2003 to Dec. 31, 2015. Realized variance (left panel) and square root of the realized quarticity (right panel).



(a) Realized variance

(b) Square root of the realized quarticity

Table 1: Five–number summary of the distributions of calculated Pearson (P) and Spearman (S) correlations across the 29 series, for various pairs of functions of (square root of) realized quarticity and realized variance.

	P	P	S	P	P	S
Vars	$RQ^{1/2}$ RV	$\ln(RQ^{1/2})$ $\ln(RV)$	$RQ^{1/2}$ RV	$RQ^{1/2}\,RV$ RV^2	$\ln(RQ^{1/2}\,RV)$ $\ln(RV^2)$	$RQ^{1/2}\,RV$ RV^2
min	0.846	0.965	0.959	0.895	0.991	0.989
Q1	0.926	0.968	0.962	0.955	0.992	0.990
median	0.938	0.972	0.966	0.975	0.992	0.991
Q3	0.954	0.974	0.969	0.987	0.993	0.992
max	0.975	0.985	0.980	0.998	0.996	0.995

The evidence from the correlation between realized quarticity and realized variance points to consistently very high values as shown in Table 1, where we reproduce the five–number summary of the distributions of such correlations across our 29 series. In the left panel, we have Pearson correlation values between $RQ^{1/2}$ and RV, and $\ln RQ^{1/2}$ and $\ln RV$, and Spearman correlations between $RQ^{1/2}$ and RV. In view of equation (2), it is instructive to see that the correlations are even higher if we consider the correlations between functions of the interaction term $RQ^{1/2}\,RV$ and RV^2.

The high correlation between $RQ^{1/2}\, RV$ and RV^2 motivates the $\mathsf{HAR2}$ alternative specification

$$
\begin{aligned}
RV_t &= \omega + (\underbrace{\alpha_D + \alpha_E RV_{t-1}}_{\alpha_{1,t-1}})RV_{t-1} + \alpha_W \overline{RV}_{t-(2:5)} + \alpha_M \overline{RV}_{t-(6:22)} + \varepsilon_t \\
&= \omega + \alpha_D RV_{t-1} + \alpha_E RV_{t-1}^2 + \alpha_W \overline{RV}_{t-(2:5)} + \alpha_M \overline{RV}_{t-(6:22)} + \varepsilon_t \qquad (3)
\end{aligned}
$$

Comparison between (2) and (3) highlights the difference in the extra nonlinear term which is inserted, with two fundamentally different interpretations across the two models:

- The Q specification is motivated by the presence of measurement errors and the need to correct the attenuation bias. *De facto*, lagged volatility observations with high quarticity are more affected by measurement errors and should receive less weight in forecasting.

- The $\mathsf{2}$ specification is more motivated by an idea that volatility dynamics is linkable to market behavior, and, as such, the model would have a more structural interpretation: bursts of volatility are shortly lived, hence, in the presence of high lagged volatility, the model would hasten the return to normality by paying less attention, so–to–speak, to those observations *irrespective of the presence of measurement errors*.

In Table 2 we report the estimates of the base HAR, the HARQ and the $\mathsf{HAR2}$ for three different choices of sampling intervals in the construction of the realized variance of SPY ETF, as a robustness check on the sensitivity of the curvature effect to the construction of the series. The results show the significance of the α_E parameters across the Q and $\mathsf{2}$ specifications, confirming the need to augment the base HAR model. An improvement in Mean Square Error (MSE) and Mean Absolute Error (MAE) is noticeable and substantially equivalent across the two augmented specifications, accompanied by the disappearance of significance of the monthly coefficient α_M. The results are consistent across sampling interval choices.

The observational equivalence between the two models HARQ and $\mathsf{HAR2}$, and the characteristic behavior relative to the HAR model are shown in Figure 2 where we mirror Figure 2 (center and right panels) of Bollerslev *et al.* (2016) for the October 6–17, 2008 highly turbulent period. The left–hand side panel shows the behavior of the time–varying coefficient $\alpha_{1,t-1}$ (including the constant level for HAR), which clearly drops in the presence of the burst of volatility on October 10 (a Friday) which reflects on the model predictions for Oct. 13; the right–hand side panel shows how the fitted behavior is modified accordingly in the augmented models, avoiding the HAR overprediction.

When the analysis is repeated across the panel of DJ30 components we get similar results, in the sense that (including the S&P500) we get 25 out of 29 cases in which α_E is significant (at 1% significance level) in the HARQ model and 26 out of 29 in the $\mathsf{HAR2}$ specification for the five-minute sampling interval. [3]

[3]Results change only slightly at different sampling frequency, being 29 and 28, respectively, in the one-minute case, and 27, respectively, 24 in the ten-minute case.

Table 2: Estimates of the HAR, HARQ and HAR2 on the realized variance of the Standard & Poor's SPY ETF. Estimation period: Jan. 3, 2003 – Dec. 31,2010. The estimates are repeated across the intra–daily sampling periods to construct the realized variance series. Newey–West HAC standard errors in parenthesis.

	5 minutes			1 minute			10 minutes		
		Q	2		Q	2		Q	2
ω	0.228	-0.167	-0.231	0.274	-0.173	-0.227	0.213	-0.228	-0.228
	(0.106)	(0.123)	(0.176)	(0.146)	(0.137)	(0.175)	(0.091)	(0.167)	(0.167)
α_D	0.331	0.856	0.843	0.285	0.931	0.895	0.367	0.851	0.879
	(0.092)	(0.106)	(0.149)	(0.119)	(0.130)	(0.158)	(0.081)	(0.157)	(0.175)
α_E		-0.382	-0.527		-0.350	-0.456		-0.447	-0.522
		(0.051)	(0.114)		(0.066)	(0.089)		(0.123)	(0.148)
α_W	0.443	0.284	0.314	0.470	0.273	0.315	0.398	0.271	0.256
	(0.148)	(0.089)	(0.093)	(0.161)	(0.096)	(0.098)	(0.132)	(0.054)	(0.065)
α_M	0.144	0.002	0.015	0.154	-0.060	-0.051	0.154	0.047	0.037
	(0.062)	(0.092)	(0.111)	(0.084)	(0.118)	(0.123)	(0.058)	(0.052)	(0.066)
MSE	21.298	18.702	19.195	27.696	23.403	23.585	19.661	17.886	17.826
MAE	1.232	1.212	1.231	1.267	1.203	1.210	1.219	1.247	1.248

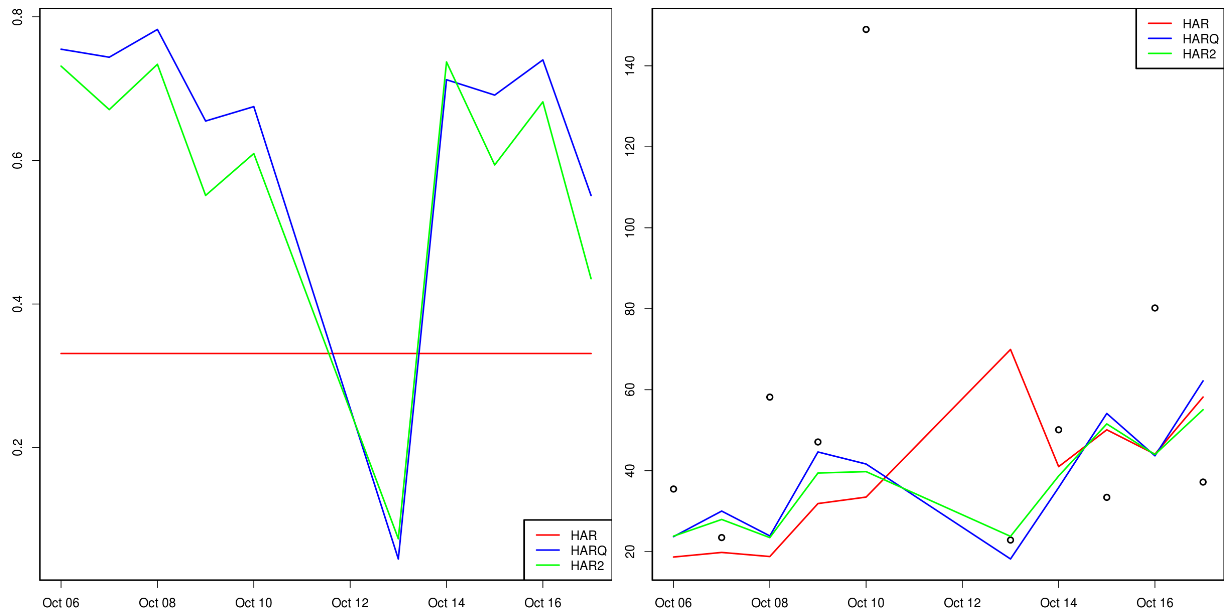
Focusing on model diagnostics, though, we find that all three specifications have substantial residual correlation. Still summarizing across the 29 variances, we find significance of the Ljung–Box joint test statistic at lag 5 for residuals, absolute residuals and squared residuals in a number of cases reported in Table 3. This finding is consistent with the findings by Corsi *et al.* (2008) about the presence of clustering in the residuals of HAR.

Table 3: Number of assets for which the Ljung–Box statistics at lag 5 turns out to be statistically significant at 1% significance level. Estimation period: Jan. 3, 2003 – Dec. 31,2010. The estimates are repeated across the intra–daily sampling periods to construct the realized variance series. LB_5 is calculated on residuals, LBa_5 on absolute residuals, and $LB2_5$ on squared residuals.

	5 minutes			1 minute			10 minutes		
	HAR	HARQ	HAR2	HAR	HARQ	HAR2	HAR	HARQ	HAR2
LB_5	27	26	26	27	27	26	26	26	25
LBa_5	29	29	29	28	28	28	29	29	29
$LB2_5$	27	22	21	20	17	16	24	22	21

This provides the two augmented models with yet another interpretation/question: are they describing appropriate volatility dynamics (be they measurement error or burst of volatility motivated) or, rather, are they linear specifications in which a interaction/curvature term is inserted to actually capture a mis-specification in the HAR model? This would be reminiscent of a White (1980)–type test where significance of squares and cross–products coefficients is used as model diagnostics. If that were the case, the indications from previous

Figure 2: SPY: Examples of estimated $\alpha_{1,t-1}$ for different HAR models (left panel) and corresponding actual values (open dots) and volatility forecasts (right panel). October 2008, cf. Fig 2 in BPQ.



(a) $\alpha_{1,t-1}$ (b) 1-step ahead forecasts

results is to explore alternative models which reduce the residual diagnostics problems and still allow to address the issue as of whether an interaction/curvature term is significant.

4 Heteroskedastic Behavior, Asymmetry and Regimes

4.1 The HAR-GARCH Models

The issue of volatility of volatility playing a role in modeling volatility was recognized in the HAR family by Corsi *et al.* (2008) who refined the features of the HAR model adding a GARCH structure on the innovation term: such a HAR-GARCH model has the advantage of accounting for clustering in the residuals, which is indeed capturing the time–varying features of volatility of volatility. In the context of the discussion here, considering a general specification which can be augmented for interaction/curvature terms, such a model can be written as

$$RV_t = \omega + (\alpha_D + \alpha_E X_{t-1})RV_{t-1} + \alpha_W \overline{RV}_{t-(2:5)} + \alpha_M \overline{RV}_{t-(6:22)} + \sqrt{h_t}u_t, \tag{4}$$

where $\varepsilon_t \equiv \sqrt{h_t}u_t$ has the usual multiplicative structure for the overall innovation under GARCH where $h_t = E(\varepsilon_t^2|\mathcal{I}_{t-1})$ provided that $u_t \sim (0,1)$, and h_t has the usual conditionally autoregressive (with variants) specification. The base model corresponds to $X_{t-1} = 0$, while the HARQ-GARCH has $X_{t-1} = RQ_{t-1}^{1/2}$ and the HAR2-GARCH has $X_{t-1} = RV_{t-1}$, the latter two being the augmented models which allow for a time–varying coefficient on RV_{t-1}, as

Table 4: Coefficient estimates on SPY realized variance for the HAR (first set of results of Table 2) and HAR-GARCH models. Sample period: Jan. 3, 2003–Dec. 31, 2010.

	HAR			HAR-GARCH		
		Q	2		Q	2
ω	0.228	-0.167	-0.231	0.100	0.096	0.099
	(0.106)	(0.123)	(0.176)	(0.034)	(0.034)	(0.034)
α_D	0.331	0.856	0.843	0.439	0.444	0.439
	(0.092)	(0.106)	(0.149)	(0.036)	(0.038)	(0.037)
α_E		-0.382	-0.527		-0.067	-0.006
		(0.051)	(0.114)		(0.086)	(0.187)
α_W	0.443	0.284	0.314	0.363	0.361	0.363
	(0.148)	(0.089)	(0.093)	(0.030)	(0.030)	(0.030)
α_M	0.144	0.002	0.015	0.103	0.103	0.103
	(0.062)	(0.092)	(0.111)	(0.037)	(0.036)	(0.036)
Const				0.039	0.029	0.029
				(0.011)	(0.011)	(0.011)
ARCH				0.236	0.234	0.236
				(0.030)	(0.030)	(0.030)
GARCH				0.763	0.765	0.763
				(0.004)	(0.004)	(0.004)
MSE	21.298	18.702	19.195	21.604	20.611	21.532
MAE	1.232	1.212	1.231	1.166	1.151	1.165

before.

In Table 4, for comparison's sake we reported the first set of results of Table 2 related to the HAR model estimated on 5–minute realized variance, next to the corresponding results for the HAR-GARCH model. After accounting for heteroskedasticity in the residuals, the striking result is the statistical insignificance of the interaction/curvature terms. Comparing the simpler versions of either model, we notice a higher relevance given to the daily term, and smaller coefficients for the weekly and monthly terms. We keep that in mind when interpreting the relative performance both in– and out–of–sample between the two models.

4.2 The AMEMs

An alternative approach to modeling realized variance is the MEM. A quite general formulation for RV_t is

$$RV_t = \mu_t \, \varepsilon_t \qquad \mu_t = \mu(\mathcal{I}_{t-1}; \theta) \qquad \varepsilon_t \overset{i.i.d.}{\sim} D^+(1, \sigma^2).$$

Here, ε_t is an i.i.d. innovation term following a generic distribution over a non–negative support with unit mean and variance σ^2. The model has the important feature (in this context) that the dependent variable is conditionally heteroskedastic:

$$E(RV_t|\mathcal{I}_{t-1}) = \mu_t \qquad V(RV_t|\mathcal{I}_{t-1}) = \sigma^2\mu_t^2$$

In what follows, we will adopt an Asymmetric MEM (AMEM) specification for μ_t (Engle and Gallo, 2006), where past values of RV_{t-1} have a different impact on the current RV_t according to the sign of past returns on that asset. Following Engle and Gallo (2006) we adopt a Gamma distribution for ε_t dependent on just one parameter a, as implied by the unit expectation constraint. We have

$$\begin{aligned}
&\mu_t = \omega + \beta_1\mu_{t-1} + (\alpha_1 + \alpha_E X_{t-1})RV_{t-1} + \gamma_1 RV_{t-1}^{(-)} \\
&\varepsilon_t|\mathcal{I}_{t-1} \sim Gamma(a, 1/a)
\end{aligned}$$

Again, the base AMEM model corresponds to $X_{t-1} = 0$, setting $X_{t-1} = RQ_{t-1}^{1/2}$ delivers the AMEMQ, while $X_{t-1} = RV_{t-1}$ corresponds to the AMEM2 , with the latter having the interaction/curvature terms, correspondingly.

4.3 The MS-AMEMs

In the augmented models the main feature is a faster mean reversion behavior: thus an additional question arises as to whether the mean reversion occurs to a constant level or to a time–varying level. To that end, among the many possible specifications which allow for a low frequency component which slowly evolves through time (cf., for example, Engle and Rangel, 2008; Brownlees and Gallo, 2010; Barigozzi *et al.*, 2014; Gallo and Otranto, 2017), we choose the Markov Switching AMEM (cf. Gallo and Otranto, 2015, for details) again allowing for all three versions, the base MS-AMEM, and the augmented ones, MS-AMEMQ and MS-AMEM2.

$$\begin{aligned}
\mu_{t,s_t} &= \omega_{s_t} + \beta_{s_t}\mu_{t-1,s_{t-1}} + (\alpha_{s_t} + \alpha_E X_{t-1})RV_{t-1} + \gamma_{s_t} RV_{t-1}^{(-)} \\
\varepsilon_t|s_t, \mathcal{I}_{t-1} &\sim Gamma(a_{s_t}, 1/a_{s_t})
\end{aligned}$$

where $s_t \in \{1, 2, 3\}$[4] and $P(s_t = j|s_{t-1} = i) = p_{ij}$ ($\sum_i p_{ij} = 1$), and X_t is defined as before.

The estimation results from the AMEM, AMEMQ, AMEM2 and MS-AMEM (with three states) are reported in Table 5: to be noticed the insignificance of the interaction and curvature effects (coefficient α_E), leading to a substantial equivalence of the three models, while the asymmetric effects play an important role from a statistical point of view. For the MS-AMEMQ and MS-AMEM2, for this sample period and this asset, the estimation converged to $\hat{\alpha}_E$ being numerically equal to zero,[5] hence we do not report the estimation results in the table.

[4]We adopt three states as they proved suitable to capture the features of the data for similar sample periods (cf. the results in Gallo and Otranto, 2015).

[5]This is not always true across assets and estimation periods, hence we can still evaluate the performance of the augmented models in terms of in– and out–of–sample prediction.

Table 5: Coefficient estimates on SPY realized variance within the AMEM and MS-AMEM groups. Sample period: Jan. 3, 2003–Dec. 31, 2010.

	AMEM	AMEMQ	AMEM2	MS(3)-AMEM
ω	0.177	0.176	0.177	0.138
	(0.008)	(0.008)	(0.008)	(0.009)
k_3				0.945
				(0.093)
α_1	0.482	0.485	0.482	0.155
	(0.033)	(0.033)	(0.033)	(0.018)
α_3				0.005
				(0.016)
α_E		-0.059	-0.0002	
		(0.082)	(0.004)	
β_1	0.332	0.330	0.332	0.562
	(0.032)	(0.032)	(0.032)	(0.025)
β_2				0.884
				(0.019)
β_3				0.875
				(0.013)
γ_1	0.189	0.190	0.189	0.128
	(0.031)	(0.032)	(0.031)	(0.021)
γ_2				0.116
				(0.019)
γ_3				0.114
				(0.027)
p_{11}				0.938
				(0.019)
p_{12}				0.049
				(0.018)
p_{21}				0.059
				(0.022)
p_{22}				0.919
				(0.020)
p_{31}				0.071
				(0.006)
p_{32}				0.078
				(0.016)
a_1	3.777	3.777	3.777	5.809
	(0.232)	(0.232)	(0.231)	(0.331)
a_2				7.213
				(0.406)
a_3				2.020
				(0.476)
MSE	22.923	21.196	22.918	19.800
MAE	1.194	1.175	1.194	1.053

5 A Simulation Study

In order to shed some light in the behavior of these models, and especially about the presence of the interaction/curvature effects, we have performed a simple simulation exercise in which the data are generated by an AMEM or by an MS-AMEM,[6] and test how many times the $\hat{\alpha}_E$ coefficient turns out to be significantly negative at two different significance levels, using both OLS standard errors and HAC (Newey and West, 1987) in estimated HAR2 models for different sample periods T.

The results, reported in Table 6, show fairly clearly that in the presence of heteroskedasticity, and even more so when also regimes are introduced, one is very likely to get a significant $\hat{\alpha_E}$ coefficient, even using HAC robust standard errors. The results on the importance of a proper inclusion of heteroskedasticity are confirmed when considering the results obtained estimating with HAR-GARCH (bottom panel of the same table). Also in this case, we distinguish between calculations of standard errors according to the inverse of the Hessian (ML), or robust (sandwich estimator). The rates of failures to reject are much lower, less so when the data are generated according to the MS-AMEM, signifying in both cases that the presence of regimes in a multiplicatively heteroskedastic model may be mistaken for the significance of curvature (or interaction) terms.

Table 6: Simulation results. Percentage of times the $\hat{\alpha}_E$ coefficient turns out to be significantly negative at different significance levels in an HAR2 model (top panel with OLS and HAC standard errors) and in an HAR2-GARCH model (bottom panel with ML and Sandwich based standard errors). Data simulated under AMEM and MS-AMEM DGPs (with estimated SPY coefficients – cf. Table 5). 1000 replications each.

Model	DGP	T	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
			OLS		HAC	
HAR2	AMEM	1000	83.0	78.8	63.9	45.8
		2000	85.8	81.9	63.8	47.6
		3000	86.0	82.9	64.3	48.0
	MS-AMEM	1000	98.5	97.3	90.5	80.5
		2000	99.1	98.4	92.2	82.2
		3000	99.2	98.6	92.9	84.1
			ML		Sandwich	
HAR2-GARCH	AMEM	1000	15.5	4.6	12.7	3.7
		2000	17.0	6.2	14.2	4.6
		3000	15.3	6.4	13.2	5.1
	MS-AMEM	1000	39.2	25.3	23.6	9.9
		2000	45.5	29.8	27.7	10.6
		3000	51.6	33.6	31.1	13.3

[6] The coefficients were set at values of the estimated for SPY – cf. Table 5. We performed 1000 replications for each combination of T and DGP. The results are similar for the HARQ model and are not reported here.

Table 7: Summary of the estimation results for the HAR-GARCH, AMEM and MS-AMEM in their base and augmented versions. In–sample period: Jan. 3, 2003–Dec. 31., 2010. Number of assets for which the estimated α_E coefficients are significant across the 29 asset realized variances, and for which the Ljung–Box statistics at lag 5 turns out to be statistically significant at 1% significance level.

	HAR-GARCH	HARQ-GARCH	HAR2-GARCH	AMEM	AMEMQ	AMEM2	MS-AMEM	MS-AMEMQ	MS-AMEM2
α_E		8	3		7	2		3	0
LB_5	7	7	9	2	2	2	9	7	7
LBa_5	5	3	4	3	3	3	2	3	3
$LB2_5$	0	0	0	0	0	0	0	0	0

6 Comparing Model Performances

6.1 In–sample

Estimated on all assets, the performance of these models is summarized in Table 7: for all 29 assets considered, we report the number of statistically significant $\hat{\alpha_E}$ coefficients, as well as the Ljung Box diagnostics (at lag 5) on residuals, absolute residuals and squared residuals. The behavior across models is fairly similar, in that we notice a much lower presence of significant $\hat{\alpha}_E$ and an overall better performance in terms of residual diagnostics.

We chose to compare the models with two loss functions, the mean absolute, respectively, squared errors. In Table 8 we report a matrix indicating the number of times the model by row has a better performance than the model by column in terms of a loss function based on MAE.[7] In parenthesis, we report the number of times that the lower loss is statistically significant according to the Diebold–Mariano test (Diebold and Mariano, 1995) at $\alpha = 0.05$. As far as HAR is concerned, it shows a relevance of the interaction/curvature term as far as the number of times turning out to be better than the base case, but rarely, if at all, statistically significant (a feature repeated throughout). All models are mostly better than HAR, and MS-AMEM has the best performance overall, significantly so in almost all cases, and within this class there is a substantial equivalence between the MS-AMEMQ and the MS-AMEM2 (in seven cases here, numerically indistinguishable).

A similar table, Table 9, reports the results for the loss function based on MSE. Being the best fit in OLS terms, it is not surprising to find a better performance of HAR (and also of HAR-GARCH) across rows, although not significantly so for many assets and models (but not for the MS-AMEM); also, the interaction/curvature terms show up as marginally important for fit. It is remarkable that also in this case MS-AMEM performs numerically better than other models in an ovewhelming majority of cases, a lot of which are also significant (roughly one out of two).

Overall these results show that, in–sample, heteroskedasticity, asymmetry, and mostly, a time–varying underlying level of average volatility capture the features of the interac-

[7]When the comparison pairs add up to less than 29, it signals a number of cases in which the loss functions coincided.

Table 8: In Sample Analysis: Absolute Errors. The table reports the number of times the model by row has a better performance than the model by column. In parenthesis, we report the number of times that the lower loss is statistically significant according to the Diebold–Mariano test (Diebold and Mariano, 1995) at $\alpha = 0.05$.

		HAR			HAR-GARCH			AMEM			MS-AMEM		
			Q	2		Q	2		Q	2		Q	2
			9(0)	9(0)	6(2)	6(2)	8(4)	9(0)	1(0)	5(0)	0	0	0
HAR	Q	20(3)		8(1)	12(2)	6(2)	11(4)	14(0)	5(0)	13(0)	0	0	0
	2	20(3)	11(1)		9(2)	7(2)	9(4)	12(0)	5(0)	11(0)	0	0	0
		23(8)	17(1)	20(0)		5(0)	13(1)	19(2)	11(0)	17(0)	0	0	0
HAR-GARCH	Q	23(16)	23(7)	22(3)	24(2)		23(1)	19(2)	17(1)	19(2)	0	0	0
	2	21(9)	18(2)	20(1)	16(0)	6(0)		18(1)	9(0)	16(0)	0	0	0
		20(0)	15(0)	17(1)	10(2)	10(2)	11(4)		1(0)	3(0)	0	0	0
AMEM	Q	28(1)	24(0)	24(1)	18(2)	12(3)	20(5)	28(4)		27(5)	0	0	0
	2	24(0)	16(0)	18(1)	12(2)	10(2)	13(4)	26(8)	2(1)		0	0	0
		29(29)	29(28)	29(28)	29(28)	29(27)	29(25)	29(27)	29(27)	29(27)		4(0)	3(0)
MS-AMEM	Q	29(29)	29(28)	29(28)	29(28)	29(27)	29(26)	29(28)	29(28)	29(28)	16(4)		13(4)
	2	29(29)	29(28)	29(28)	29(28)	29(27)	29(26)	29(28)	29(28)	29(28)	17(5)	9(3)	

tion/curvature terms inserted in the HAR models.

6.2 Out–of–sample

The model comparison is performed similarly for the out–of–sample forecast performance. As mentioned before, we estimate models starting from 2003-2010 and produce one–step ahead forecasts for the following year; we first move the sample period ahead by one year (i.e. 2004–2011) and we forecast the year 2012, repating the procedure until we estimate over 2007-2014 and we forecast over 2015.

With the forecast errors thus derived we build two tables, similar to what we encountered before. Table 10, based on absolute errors we have some interesting differences relative to the in–sample analysis. First, the augmented HAR models perform better than the HAR, confirming the need for the extra term, be it the quarticity or the realized variance itself (significant cases increase, while confirming a balance between the Q and the 2 cases). The lack of significant contribution by the same terms is confirmed in all other cases: if we look at the boxes on the diagonals, we hardly find significant differences relative to the base model. A novelty is the strong performance of the HAR-GARCH with a strong number of assets giving better results relative to the AMEM. We confirm the overall dominance of the MS-AMEM over all other models with almost all assets showing significantly better performances than with the other models.

Qualitatively, the same remarks apply to the MSE based loss function, with the strongest performance recorded for the MS-AMEMs (with no difference between the augmented and the base models), with a slightly lower number of significant cases across.

Table 9: In Sample Analysis: Squared Errors. The table reports the number of times the model by row has a better performance than the model by column. In parentheses, we report the number of times that the lower loss is statistically significant according to the Diebold–Mariano test (Diebold and Mariano, 1995) at $\alpha = 0.05$.

		HAR			HAR-GARCH			AMEM			MS-AMEM		
			Q	2		Q	2		Q	2		Q	2
			0	1(0)	29(2)	16(2)	20(2)	25(0)	13(0)	23(0)	1(0)	0	0
HAR	Q	29(1)		17(0)	29(2)	29(1)	29(1)	29(0)	27(0)	28(0)	6(0)	5(0)	5(0)
	2	28(0)	12(0)		29(1)	29(1)	29(1)	29(0)	28(0)	29(0)	5(0)	3(0)	2(0)
		0	0	0		7(0)	12(0)	16(0)	6(0)	13(0)	1(0)	0	0
HAR-GARCH	Q	13(0)	0	0	22(0)		18(0)	18(0)	11(0)	16(0)	2(0)	1(0)	1(0)
	2	9(0)	0	0	17(0)	11(0)		14(0)	7(0)	13(0)	2(0)	1(0)	1(0)
		4(0)	0	0	13(0)	11(0)	15(0)		1(0)	4(0)	1(0)	0	0
AMEM	Q	16(0)	2(0)	1(0)	23(0)	18(0)	22(1)	28(0)		24(0)	3(0)	2(0)	2(0)
	2	6(0)	1(0)	0	16(0)	13(0)	16(0)	25(1)	5(0)		1(0)	0	0
		28(15)	23(11)	24(15)	28(12)	27(14)	27(12)	28(12)	26(15)	28(13)		10(1)	8(2)
MS-AMEM	Q	29(16)	24(15)	26(17)	29(11)	28(15)	28(13)	29(13)	27(16)	29(14)	10(4)		9(2)
	2	29(15)	24(13)	27(17)	29(12)	28(15)	28(14)	29(13)	27(16)	29(14)	12(5)	13(4)	

7 Concluding Remarks

Dynamic modeling of realized variance behavior is important to forecast conditional return variance, taking full advantage of the better properties of the ultra–high frequency–based estimates of variance. These being estimates they are affected by measurement errors, as discussed in all of the realized variance literature recalled in BPQ. The variability of the variance estimates crucially depends on the quarticity: in essence we know that realized variance estimates are less accurate when quarticity is high. In this paper, we stressed that the (square root of) quarticity has a high correlation with realized variance itself: this means that realized variance estimates are less accurate when the variance itself is high.

The approach suggested by BPQ emphasizes the presence of these measurement errors and the important consequence of requiring a refinement in modeling. They take the HAR model as a benchmark of realized variance modeling and insert a term which is intended to capture a bias attenuation effect in coefficient estimation of the lagged realized variance. The ensuing curvature effect (i.e. a stronger mean reversion when quarticity is high) is shown in a variety of instances to be statistically significant: in a very practical sense, the implication is that the augmented model pays smaller attention, so to speak, to lagged values of realized variance when quarticity is high.

In this paper we replicate BPQ's results, but we offer an alternative explanation for the curvature effect, starting from the consideration that quarticity and realized variance (under a number of transformations or correlation measures) are strongly correlated. In our view, the measurement error explanation is observationally equivalent to a more structural explanation in which the dynamics of volatility must take into account the empirical fact that burst of volatility die out more quickly, i.e. they mean–revert more strongly than other variance increases of lesser importance. The two alternative explanations were tested by considering on the one hand an *interaction term* given by the product of the square root of the quarticity times the realized variance, and, on the other, a *curvature term*, the square

Table 10: Out-of-Sample Analysis: Absolute Errors. The table reports the number of times the model by row has a better performance than the model by column. In parentheses, we report the number of times that the lower loss is statistically significant according to the Diebold–Mariano test (Diebold and Mariano, 1995) at $\alpha = 0.05$.

		HAR			HAR-GARCH			AMEM			MS-AMEM		
			Q	2		Q	2		Q	2		Q	2
			4(1)	6(2)	2(1)	2(1)	2(1)	12(1)	8(0)	12(1)	1(0)	0	1(0)
HAR	Q	25(15)		14(9)	6(1)	2(1)	2(1)	19(8)	16(7)	20(8)	1(0)	0	1(0)
	2	23(12)	15(10)		5(1)	2(1)	2(1)	18(9)	17(9)	18(9)	1(0)	0	1(0)
		27(25)	23(18)	24(21)		12(3)	12(3)	27(20)	26(16)	27(20)	1(0)	0	1(0)
HAR-GARCH	Q	27(25)	27(17)	27(19)	17(5)		0	27(21)	26(20)	27(21)	1(0)	0	1(0)
	2	27(25)	27(17)	27(19)	17(5)	0		27(21)	26(20)	27(21)	1(0)	0	1(0)
		17(8)	10(3)	11(2)	2(1)	2(1)	2(1)		2(0)	13(6)	1(0)	0	1(0)
AMEM	Q	21(8)	13(3)	12(5)	3(1)	3(1)	3(1)	27(1)		24(0)	1(0)	0	1(0)
	2	17(8)	9(3)	11(2)	2(1)	2(1)	2(1)	16(2)	5(0)		1(0)	0	1(0)
		28(28)	28(27)	28(28)	28(26)	28(27)	28(27)	28(28)	28(27)	28(28)		12(1)	11(0)
MS-AMEM	Q	29(29)	29(28)	29(29)	29(26)	29(26)	29(26)	29(28)	29(28)	29(28)	17(5)		20(8)
	2	28(28)	28(27)	28(28)	28(26)	28(27)	28(27)	28(28)	28(27)	28(28)	18(6)	9(3)	

of realized variance. The two terms being very close, it is not surprising to find that they deliver very similar results within the HAR model, both in terms of numerical values and in terms of statistical significance across a representative sample of US market assets (the S&P500 plus 28 components of the Dow Jones index). The results are robust to the choice of sampling frequency (1–, 5–, and 10–minutes) in constructing the realized variance.

A second, perhaps, more fundamental issue arises when the residuals of the base HAR model and of its two augmented versions (namely, the HARQ and the HAR2) are subject to diagnostics (autocorrelation in the residuals, in the absolute residuals and in the squared residuals). All statistics show signs of misspecification of this class of HAR models providing the possible explanation that the significance of the interaction/curvature terms signals the need for a different specification. This is reminiscent of a specification test *à la* White (1980) which shows significant nonlinear elements, but is not constructive as what direction to take to improve upon the extant choice.

We have explored several directions, all pointing to the importance of adopting a heteroskedastic model which gives different importance to observations: the HAR-GARCH model of Corsi *et al.* (2008) where HAR is enriched by a GARCH–type structure to account for the volatility of volatility (related to quarticity), the AMEM model of Engle and Gallo (2006), where a multiplicative error structure makes the variance of realized variance proportional to the square of its expectation, and the MS-AMEM model of Gallo and Otranto (2015) where, in addition, different regimes are considered, and dynamics is regulated by regime–specific coefficients (both the speed of mean reversion, and, empirically most importantly, the level of the mean to which the model reverts change). In the latter two, the reaction of realized variance to the sign of past returns turns out to be an important driver of the dynamics which is not present in the other models.

In all these models the statistical significance of the interaction/curvature terms is almost always disappearing (more strongly so for the MS-AMEM), confirming the view that

Table 11: Out-of-Sample Analysis: Squared Errors. The table reports the number of times the model by row has a better performance than the model by column. In parentheses, we report the number of times that the lower loss is statistically significant according to the Diebold–Mariano test (Diebold and Mariano, 1995) at $\alpha = 0.05$.

		HAR			HAR-GARCH			AMEM			MS-AMEM		
			Q	2		Q	2		Q	2		Q	2
HAR			15(1)	21(0)	7(0)	9(0)	9(0)	15(0)	7(0)	15(0)	0	0	0
	Q	14(1)		18(2)	8(1)	8(0)	8(0)	12(0)	5(0)	12(0)	0	0	1(0)
	2	8(0)	11(2)		5(1)	6(0)	6(0)	12(0)	6(0)	12(0)	0	0	0
HAR-GARCH		22(5)	21(3)	24(0)		10(1)	10(1)	19(1)	14(1)	20(1)	0	0	0
	Q	20(8)	21(3)	23(1)	19(0)		0	19(1)	14(1)	19(1)	0	0	1(0)
	2	20(8)	21(3)	23(1)	19(0)	0		19(1)	14(1)	19(1)	0	0	1(0)
AMEM		14(0)	17(6)	17(7)	10(1)	10(0)	10(0)		1(0)	15(3)	0	0	0
	Q	22(0)	24(12)	23(10)	15(1)	15(1)	15(1)	28(0)		27(0)	0	0	0
	2	14(0)	17(6)	17(7)	9(1)	10(0)	10(0)	14(0)	2(0)		0	0	0
MS-AMEM		29(22)	29(19)	29(18)	29(19)	29(16)	29(16)	29(14)	29(18)	29(15)		15(0)	14(0)
	Q	29(22)	29(20)	29(19)	29(21)	29(20)	29(20)	29(15)	29(18)	29(16)	14(2)		17(0)
	2	29(22)	28(19)	29(19)	29(21)	28(19)	28(19)	29(15)	29(19)	29(16)	15(3)	12(0)	

the HAR class of models for variance is misspecified if one does not take heteroskedasticity and the presence of regime–specific dynamics into consideration. This result is strengthened by a simulation study where we generate observations according to the AMEM and the MS-AMEM (without interaction/curvature terms) and when HAR is estimated with either of those terms, it turns up being statistically significant.

The analysis we performed both in– and out–of–sample shows that the MS-AMEM is by far the best performing model, with a good second–best performance of the HAR-GARCH model (even over AMEM). The augmented versions of the models show no sign of improved performance in these cases, while such better performance is confirmed for HARQ and HAR2 over HAR.

We are certainly not intending to deny the documented presence of measurement errors in estimating realized variance. We, rather, want to provide a strong word of a caution away from interpreting some features of misspecification as a sign of measurement errors biasing coefficient in a relevant way. Heteroskedasticity (volatility of volatility) and a time-varying mean seem to play a role of higher order of importance than measurement errors. The important message contained in BPQ is that measurement errors are more important with high (sqrt–) quarticity and that should be corrected for. We show that the latter is highly correlated with realized variance itself, and hence, in general, we should pay less attention to high values of variance when modeling its dynamics for prediction purposes. Heteroskedastic models, such as HAR-GARCH and AMEM are equipped to do so, the latter from the outset and not as a patch. Moreover, we think that the issue of a low frequency volatility component being present in the data is unrelated to measurement errors and needs to be adequately addressed.

All analysis here is conducted for the realized variance as in BPQ. In the HAR model this is a natural choice because of the additivity of the variance when aggregating at the weekly and monthly levels. An alternative, not analyzed here, is to model volatility directly, rather than variance as in Cipollini *et al.* (2013) and Gallo and Otranto (2015), in view of the fact that the volatility has less severe peaks in correspondence to volatility bursts. As a matter of fact, Cipollini *et al.* (2013) show that there is a definite improvement in the forecasting performance of the (multivariate) modeling of realized volatility over realized variance.

References

Alizadeh, S., Brandt, M. W. and Diebold, F. X. (2002) Range-based estimation of stochastic volatility models, *The Journal of Finance*, **57**, 1047–1091.

Amado, C. and Teräsvirta, T. (2014) Modelling changes in the unconditional variance of long stock return series, *Journal of Empirical Finance*, **25**, 15–35.

Andersen, T. G. and Benzoni, L. (2009) Realized volatility, in *Handbook of Financial Time Series* (Eds.) T. G. Andersen, R. A. Davis, J. P. Kreiss and T. Mikosch, Springer Verlag.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F. and Diebold, F. X. (2006) Volatility and correlation forecasting, in *Handbook of Economic Forecasting* (Eds.) G. Elliott, C. W. J. Granger and A. Timmermann, North Holland.

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, *Review of Economics and Statistics*, **89**, 701–720.

Barigozzi, M., Brownlees, C., Gallo, G. and Veredas, D. (2014) Disentangling systematic and idiosyncratic dynamics in panels of volatility measures, *Journal of Econometrics*, **182**, 364–384.

Barndorff-Nielsen, O. 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)*, **64**, 253–280.

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

Brownlees, C. T. and Gallo, G. M. (2006) Financial econometric analysis at ultra–high frequency: Data handling concerns, *Computational Statistics and Data Analysis*, **51**, 2232–2245.

Brownlees, C. T. and Gallo, G. M. (2010) Comparison of volatility measures: a risk management perspective, *Journal of Financial Econometrics*, **8**, 29–56.

Brownlees, C. T. and Gallo, G. M. (2011) Shrinkage estimation of semiparametric multiplicative error models, *International Journal of Forecasting*, **27**, 365–378.

Cipollini, F., Engle, R. F. and Gallo, G. M. (2013) Semiparametric vector MEM, *Journal of Applied Econometrics*, **28**, 1067–1086.

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

Corsi, F., Mittnik, S., Pigorsch, C. and Pigorsch, U. (2008) The volatility of realized volatility, *Econometric Reviews*, **27**, 46–78.

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

Dueker, M. J. (1997) Markov switching in garch processes and mean-reverting stock-market volatility, *Journal of Business & Economic Statistics*, **15**, 26–34.

Engle, R. F. (2002) New frontiers for ARCH models, *Journal of Applied Econometrics*, **17**, 425–446.

Engle, R. F. and Gallo, G. M. (2006) A multiple indicators model for volatility using intra-daily data, *Journal of Econometrics*, **131**, 3–27.

Engle, R. F. and Rangel, J. G. (2008) The spline-GARCH model for low frequency volatility and its global macroeconomic causes, *Review of Financial Studies*, **21**, 1187–1222.

Gallo, G. M. and Otranto, E. (2015) Forecasting realized volatility with changing average levels, *International Journal of Forecasting*, **31**, 620–634.

Gallo, G. M. and Otranto, E. (2017) Combining sharp and smooth transitions in volatility dynamics: a fuzzy regime approach, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, forthcoming.

Haas, M., Mittnik, S. and Paolella, M. (2004) A new approach to Markov-switching GARCH models, *Journal of Financial Econometrics*, **2**, 493 – 530.

Hansen, P. R. (2010) A winner's curse for econometric models: on the joint distribution of in-sample fit and out-of-sample fit and its implications for model selection, Tech. rep., Stanford University.

Maheu, J. M. and McCurdy, T. H. (2002) Nonlinear features of realized FX volatility, *Review of Economics and Statistics*, **84**, 668–681.

McAleer, M. and Medeiros, M. (2008) A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries., *Journal of Econometrics*, **147**, 104–119.

Newey, W. and West, K. (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, **55**, 703–08.

Teräsvirta, T. (2009) An introduction to univariate GARCH models, in *Handbook of Financial Time Series* (Eds.) T. G. Andersen, R. A. Davis, J. P. Kreiss and T. Mikosch, Springer Verlag, chap. 1, pp. 17–42.

Whaley, R. E. (2009) Understanding the VIX, *The Journal of Portfolio Management*, **35**, 98105.

White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, **48**, 817–38.

Appendix – List of tickers

Table 12: List of tickers of the realized variance series

AAPL	Apple
AXP	American Express
BA	Boeing
CAT	Caterpillar
CSCO	Cisco
CVX	Chevron
DD	DuPont
DIS	Disney
GE	General Electric
GS	Goldman Sachs
HD	Home Depot
IBM	IBM
INTC	Intel
JNJ	Johnson and Johnson
JPM	J.P. Morgan Chase
KO	Coca Cola
MCD	McDonald's
MMM	3M
MRK	Merck
MSFT	Microsoft
NKE	Nike
PFE	Pfizer
PG	Procter and Gamble
SPY	SPY ETF
UNH	United Health
UTX	United Technologies
VZ	Verizon
WMT	Walmart
XOM	Exxon Mobil