Studies in Nonlinear Dynamics & Econometrics

Volume 9, *Issue* 4 2005 *Article* 6

Forecasting Stock Market Volatility with Regime-Switching GARCH Models

Juri Marcucci*

*University of California, San Diego, juri@sssup.it

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, bepress, which has been given certain exclusive rights by the author. *Studies in Nonlinear Dynamics & Econometrics* is

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Forecasting Stock Market Volatility with Regime-Switching GARCH Models*

Juri Marcucci

Abstract

In this paper we compare a set of different standard GARCH models with a group of Markov Regime-Switching GARCH (MRS-GARCH) in terms of their ability to forecast the US stock market volatility at horizons that range from one day to one month. To take into account the excessive persistence usually found in GARCH models that implies too smooth and too high volatility forecasts, in the MRS-GARCH models all parameters switch between a low and a high volatility regime. Both gaussian and fat-tailed conditional distributions for the residuals are assumed, and the degrees of freedom can also be state-dependent to capture possible time-varying kurtosis. The forecasting performances of the competing models are evaluated both with statistical and risk-management loss functions. Under statistical losses, we use both tests of equal predictive ability of the Diebold-Mariano-type and test of superior predictive ability. Under risk-management losses, we use a two-step selection procedure where we first check which models pass the tests of correct unconditional or conditional coverage and then we compare the best models under two subjective VaR-based loss functions. The empirical analysis demonstrates that MRS-GARCH models do really outperform all standard GARCH models in forecasting volatility at horizons shorter than one week under both statistical and VaR-based risk-management loss functions. In particular, all tests reject the presence of a better model than the MRS-GARCH with normal innovations. However, at forecast horizons longer than one week, standard asymmetric GARCH models tend to be superior.

*The previous version of this paper has been awarded the "SNDE Best Graduate Student Paper Prize" at the Eleventh Annual Symposium of the Society for Nonlinear Dynamics and Econometrics, held in Florence, March 2003. I would like to thank the editor and an anonymous referee for helpful and constructive comments that greatly helped me improving the paper. I also would like to thank Carlo Bianchi, Graham Elliott, Robert Engle, Giampiero Gallo, Raffaella Giacomini, Clive Granger, James Hamilton, Bruce Lehmann, Francesca Lotti, Andrew Patton, Kevin Sheppard, Allan Timmermann, Halbert White and all the participants in the Symposium for valuable and helpful comments. All errors remain, of course, my own responsibility. Contact information: Juri Marcucci, Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla CA 92093-0508, USA. E-mail: jmarcucc@weber.ucsd.edu

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

1 Introduction

In the last few decades a growing number of scholars have focused their attention on modeling and forecasting volatility, due to its crucial role in financial markets. Portfolio managers, option traders and market makers are interested in the possibility of forecasting, with a reasonable level of accuracy, this important magnitude, since it is a key ingredient for portfolio optimization, derivative pricing and Value-at-Risk (VaR).

So far in the literature, many volatility models have been put forward, but those that seem to be the most successful are the GARCH (Generalized Autoregressive Conditional Heterskedasticity) models by Bollerslev (1986), who generalizes the seminal idea on ARCH by Engle (1982), and their numerous generalizations that add asymmetries, long memory, or structural breaks. GARCH models are incredibly popular because of their ability to capture many of the typical stylized facts of financial time series, such as time-varying volatility, persistence and volatility clustering.

Andersen and Bollerslev (1998) find that GARCH models do really provide good volatility forecasts, in particular when a good proxy for the latent volatility, such as the realized volatility, is adopted. Conversely, when a lousy measure for the ex-post volatility, such as the squared returns, is used, GARCH models tend to give a good in-sample fit, but very poor forecasting performances.

One of the main goals of the present paper is to show that possible concerns about the forecasting ability of standard GARCH models can also arise because of structural breaks that may lead to the usually too high estimated persistence of the individual volatility shocks (see Lamoureux and Lastrapes, 1990). Hamilton and Susmel (1994), for example, find that, for their weekly stock return data, a shock on a given week would produce non-negligible effects on the variance more than one year later. This can be one of the main reasons why GARCH volatility forecasts are sometimes too smooth and too high across periods with different levels of turbulence. Financial returns exhibit sudden jumps due not only to structural breaks in the real economy, but also to changes in the operators' expectations about the future, stemming from different information or dissimilar preferences. The real volatility is affected by million of shocks, that never persist for a long time, rendering its behavior mean-reverting. It follows that, to give better forecasts, a good volatility model should entail a different way of treating shocks. For these reasons, in the present work, we focus on the forecasting ability of Markov Regime-Switching GARCH (MRS-GARCH) models where a GARCH model is incorporated into a regime-switching framework, that allows, rather par-

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

simoniously, to take into account the existence of different volatility regimes. In particular, we consider two volatility regimes characterized by a low and a high level of turbulence. In both regimes, volatility follows a GARCH-like pattern, in such a way to avoid path-dependence as in Klaassen (2002).

The literature on MRS-GARCH models begins with Cai (1994) and Hamilton and Susmel (1994) who are the first to apply simultaneously the seminal idea of endogenous regime-switching parameters by Hamilton (1988, 1989 and 1990) into an ARCH specification to account for the possible presence of structural breaks. Nevertheless, they use an ARCH specification instead of a GARCH to overcome the problem of infinite path-dependence, i.e. to avoid the conditional variance at time t depending on the entire sample path. Gray (1996) suggests to integrate out the unobserved regime path in the GARCH equation using the conditional expectation of the past variance and his model can be regarded as the first MRS-GARCH. Recently, Klaassen (2002) suggests adopting the conditional expectation of the lagged conditional variance with a broader information set than in Gray (1996), so that his MRS-GARCH model has two main advantages over Gray's (1996) framework. It allows both a higher flexibility in capturing the persistence of volatility shocks, and gives straightforward expressions for multi-step-ahead volatility forecasts that can be calculated recursively as in standard GARCH models. Recently, Haas *et al.* (2004a) have introduced a new MRS-GARCH model that generalizes the previous MRS-GARCH models to a multi-regime setting which has the advantage of being analytically tractable and allows the authors to derive richer dynamics for the process involved. This model has two special cases: the first one is the Markov-switching ARCH as in Franq *et al.* (2001), while the second is a mixed-normal conditional heteroskedastic model as in Haas *et al.* (2004b), where mixed-normal distributions are coupled with a GARCH-type structure to resemble the features of a MRS-GARCH. Calvet and Fisher (2004) introduce a Markov Switching Multifractal model which is compared to a standard GARCH and a MRS-GARCH but, nevertheless, their model does not have a real GARCH structure in it.

In the present paper we adopt Klaassen's (2002) MRS-GARCH model using both normal and fat-tailed distributions, such as the Student's t and the GED. In addition, the shape parameter is allowed to vary across different regimes, to capture possible time-varying conditional kurtosis in a way that generalizes Dueker (1997)'s Regime-Switching GARCH models, where only few parameters are state-dependent. All these MRS-GARCH models are compared with a set of standard GARCH models in terms of their ability to forecast volatility. The forecasting performances of the competing models are evaluated for different horizons (from one

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

day to one month) and with both statistical and risk-management loss functions. Under statistical losses of the Mean Squared Error (MSE) type, the out-of-sample comparison is made both through tests of equal predictive ability (EPA) of the Diebold-Mariano-type and through tests of superior predictive ability (SPA), such as the White's (2000) Reality Check test and Hansen's (2005) SPA test.

Since volatility is used as a key ingredient for VaR estimates, a set of risk-management VaR-based loss functions is also adopted to compare the forecasting performances of the competing models. Under this VaR-based losses, we adopt the two-step procedure suggested by Sarma *et al.* (2003). In the first stage, the best models are selected according to the correct unconditional or conditional coverage of their VaR, whereas in the second stage pairs of models that pass the first-step tests are compared through the significance of their loss differentials according to two subjective economic losses that are functions of the magnitude of the VaR failures. To our knowledge, previous papers only compare different models in terms of a few statistical loss functions and, sometimes, in terms of correct unconditional coverage and, nevertheless, the comparison is in most cases carried out only with tests of EPA. In this paper, instead, we give a complete out-of-sample evaluation of the performances of a number of MRS-GARCH models and some standard GARCH.

The main empirical results, using US stock market data, point out that MRS-GARCH models significantly outperform usual GARCH in forecasting volatility at shorter horizons, while at longer ones, standard asymmetric GARCH fare better. However, none of the models seems to be uniformly superior in forecasting the US stock market volatility. Accounting for regime shifts in all the parameters of the first two moments of the conditional distribution, along with the inclusion of GARCH effects, we obtain outstanding out-of-sample results in terms of both EPA and SPA under the usual statistical loss functions of the MSE-type. In particular, at shorter horizons of one day and one week, the MRS-GARCH with normal innovations, that already outperforms all the other models under statistical losses, also fares the best in terms of a more realistic risk-management loss.

These findings partially agree with Dacco and Satchell's (1999) arguments that the choice of the correct loss function is crucial to evaluate the accuracy of volatility forecasts from non-linear models. However, our results do not confirm their main theoretical findings that although most non-linear techniques give a good in-sample fit, they are usually outperformed out-of-sample by simpler models using economic loss functions. Even though our results from the out-of-sample evaluation in terms of risk-management losses do not give a clear-cut answer, we have some evidence that MRS-GARCH models do fare better than other simpler

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

competitors, even when we use more realistic economic loss functions.

The present paper shows three main results. First, at short horizons (from one day to one week) regime-switching models with GARCH effects do really outperform standard GARCH in predicting volatility under both statistical and risk-management loss functions. Second, at horizons longer than one week, standard asymmetric GARCH models fare better than MRS-GARCH using both statistical and VaR-based losses. Third, under VaR-based risk-management loss functions, taking into account the opportunity cost of capital along with the magnitude of the VaR failures is crucial to select the best model.

The remaining of the paper is organized as follows. Standard GARCH models are presented in section 2 while section 3 is devoted to a detailed description of MRS-GARCH models. The stock market data (daily and intra-daily) used and the methodology to calculate the proxies for volatility are discussed in section 4. In section 5 the statistical and risk-management VaR-based loss functions are presented along with the out-of-sample tests of EPA and SPA adopted to evaluate multi-step-ahead volatility forecasts from different models. The in-sample and out-of-sample empirical results are discussed in section 6. Section 7 provides some conclusions and directions for further research.

2 GARCH Models

Let us consider a stock market index p_t and its corresponding rate of return r_t, defined as the continuously compounded rate of return (in percent)

$$r_t = 100 \left[\log \left(p_t \right) - \log \left(p_{t-1} \right) \right] \tag{2.1}$$

where the index t denotes the daily closing observations and $t = -R+1, \ldots, n$. The sample period consists of an estimation (or in-sample) period with R observations ($t = -R+1, \ldots, 0$), and an evaluation (or out-of-sample) period with n observations ($t = 1, \ldots, n$).

The GARCH(1,1) model for the series of returns r_t can be written as

$$r_t = \delta + \varepsilon_t = \delta + \eta_t \sqrt{h_t} \tag{2.2}$$

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta h_{t-1} \tag{2.3}$$

where $\alpha_0 > 0$, $\alpha_1 \geq 0$ and $\beta_1 \geq 0$ to ensure a positive conditional variance, and the innovation is conveniently expressed as the product of an *i.i.d.* process with

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

zero mean and unit variance (η_t) times the square root of the conditional variance.

In order to cope with the skewness often encountered in financial returns, Nelson (1991) introduces the Exponential GARCH (EGARCH) model where the logarithm of the conditional variance is modeled as

$$\log(h_t) = \alpha_0 + \alpha_1 \left| \frac{\varepsilon_{t-1}}{h_{t-1}} \right| + \xi \frac{\varepsilon_{t-1}}{h_{t-1}} + \beta_1 \log(h_{t-1}) \tag{2.4}$$

with no parameter constraints.

Glosten *et al.* (1993) put forward a modified GARCH model (GJR) to account for the 'leverage effect'. This is an asymmetric GARCH model that allows the conditional variance to respond differently to shocks of either sign and is defined as follows

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \left[1 - \mathcal{I}_{\{\varepsilon_{t-1}>0\}}\right] + \xi \varepsilon_{t-1}^2 \mathcal{I}_{\{\varepsilon_{t-1}>0\}} + \beta_1 h_{t-1} \tag{2.5}$$

where $\mathcal{I}_{\{\omega\}}$ is the indicator function which is equal to one when ω is true and zero otherwise.

Another common finding in the GARCH literature is the leptokurtosis of the empirical distribution of financial returns. To model such fat-tailed distributions researchers have adopted the Student's t or the Generalized Error Distribution (GED). Therefore, in addition to the classic gaussian assumption, in what follows, the errors ε_t are also assumed to be distributed according to these distributions. If a Student's t distribution with ν degrees of freedom is assumed, the probability density function (pdf) of ε_t takes the form

$$f(\varepsilon_t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)} (\nu-2)^{-\frac{1}{2}} (h_t)^{-\frac{1}{2}} \left[1 + \frac{\varepsilon_t^2}{h_t(\nu-2)}\right]^{-\frac{(\nu+1)}{2}} \tag{2.6}$$

where $\Gamma(\cdot)$ is the Gamma function and ν is the degree-of-freedom (or shape) parameter, constrained to be greater than two so that the second moments exist. Instead, with a GED distribution the pdf of the innovations becomes

$$f(\varepsilon_t) = \frac{\nu \exp\left[-\left(\frac{1}{2}\right) \left| \frac{\varepsilon_t}{\lambda h_t^{1/2}} \right|^{\nu}\right]}{h_t^{1/2} \lambda 2^{\left(1+\frac{1}{\nu}\right)} \Gamma\left(\frac{1}{\nu}\right)} \tag{2.7}$$

with $\lambda \equiv [(2^{-2/\nu}\Gamma(1/\nu))/\Gamma(3/\nu)]^{1/2}$, where $\Gamma(\cdot)$ is the Gamma function, ν is the thickness-of-tail (or shape) parameter, satisfying the condition $0 < \nu < \infty$ and indicating how thick the tails of the distribution are, compared to the normal.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

When the shape parameter $\nu = 2$, the GED becomes a standard normal distribution, while for $\nu < 2$ and $\nu > 2$ the distribution has thicker and thinner tails than the normal, respectively.

3 Markov Regime-Switching GARCH Models

The main feature of regime-switching models is the possibility for some, or all, the parameters to switch across different regimes (or *states* of the world) according to a Markov process, governed by a state variable, denoted s_t. The logic behind is having a mixture of distributions with different characteristics, from which the model draws the current value of the variable, according to the more likely (unobserved) state that could have determined its value. The state variable is assumed to evolve according to a first-order Markov chain, with transition probability

$$\Pr\left(s_t = j | s_{t-1} = i\right) = p_{ij} \tag{3.1}$$

that indicates the probability of switching from state i at time $t-1$ into state j at t. Usually these probabilities are grouped together into the transition matrix

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{bmatrix} = \begin{bmatrix} p & (1-q) \\ (1-p) & q \end{bmatrix} \tag{3.2}$$

where, for simplicity, the existence of only two regimes has been considered. The ergodic probability (that is the unconditional probability) of being in state[1] $s_t = 1$ is given by $\pi_1 = (1-q)/(2-p-q)$.

The MRS-GARCH model in its most general form can be written as

$$r_t | \zeta_{t-1} \sim \begin{cases} f\left(\theta_t^{(1)}\right) & \text{w. p. } p_{1,t} \\ f\left(\theta_t^{(2)}\right) & \text{w. p. } (1-p_{1,t}) \end{cases} \tag{3.3}$$

where $f(\cdot)$ represents one of the possible conditional distributions that can be assumed, that is Normal (N), Student's t or GED, $\theta_t^{(i)}$ denotes the vector of parameters in the i-th regime that characterize the distribution, $p_{1,t} = \Pr\left[s_t = 1 | \zeta_{t-1}\right]$ is the ex-ante probability[2] and ζ_{t-1} denotes the information set at time $t-1$, i.e the

[1] For further details on regime-switching models, see Hamilton (1994).

[2] The ex-ante probability should be denoted $p_{1,t|t-1}$ but to keep the notation simple we suppress the conditioning to ζ_{t-1}.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

σ-algebra induced by all the variables observed at $t-1$. More specifically, the vector of time-varying parameters can be decomposed as follows

$$\theta_t^{(i)} = \left(\mu_t^{(i)}, h_t^{(i)}, \nu_t^{(i)}\right) \tag{3.4}$$

where $\mu_t^{(i)} \equiv E\left(r_t|\zeta_{t-1}\right)$ is the conditional mean (or location parameter), $h_t^{(i)} \equiv Var\left(r_t|\zeta_{t-1}\right)$ is the conditional variance (or scale parameter), and $\nu_t^{(i)}$ is the shape parameter of the conditional distribution.[3] Hence, the family of density functions of r_t is a location-scale family with time-varying shape parameters in the most general setting.

Therefore, the MRS-GARCH consists of four elements: the conditional mean, the conditional variance, the regime process and the conditional distribution. The conditional mean equation, which is generally modeled through a random walk with or without drift, here is simply modeled as

$$r_t = \mu_t^{(i)} + \varepsilon_t = \delta^{(i)} + \varepsilon_t \tag{3.5}$$

where $i = 1, 2$ and $\varepsilon_t = \eta_t \left[h_t^{(i)}\right]^{1/2}$ and η_t is a zero mean, unit variance process. The main reason for this choice is our focus on volatility forecasting.

The conditional variance of r_t, given the whole regime path (not observed by the econometrician) $\widetilde{s}_t = (s_t, s_{t-1}, \ldots)$, is[4] $h_t^{(i)} = V\left[\varepsilon_t|\widetilde{s}_t, \zeta_{t-1}\right]$. For this conditional variance the following GARCH(1,1)-like expression is assumed

$$h_t^{(i)} = \alpha_0^{(i)} + \alpha_1^{(i)}\varepsilon_{t-1}^2 + \beta_1^{(i)}h_{t-1} \tag{3.6}$$

where h_{t-1} is a state-independent average of past conditional variances. Actually, in a regime-switching context, a GARCH model with a state-dependent past conditional variance would be infeasible. The conditional variance would in fact depend not only on the observable information ζ_{t-1} and on the current regime s_t which determines all the parameters, but also on all past states $\widetilde{s}_{t-1}$. This would require the integration over a number of (unobserved) regime paths that would grow exponentially with the sample size, rendering the model essentially intractable and impossible to estimate.

Therefore, a simplification is needed to avoid the conditional variance being a function of all past states. Cai (1994) and Hamilton and Susmel (1994) are the first to point out this difficulty by combining the regime-switching approach with

[3] In all formulae the superscript (i) denotes the regime in which the process is at time t.

[4] Here we use Klaassen's (2002) model, simplifying his notation.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

ARCH models only, thus eliminating the GARCH term in (3.6). However, both Cai (1994) and Hamilton and Susmel (1994) realize that many lags are needed for such processes to be sensible.

To avoid the path-dependence problem, Gray (1996) suggests to integrate out the unobserved regime path $\widetilde{s}_{t-1}$ in the GARCH term in (3.6) by using the conditional expectation of the past variance. In particular, Gray (1996) uses the information observable at time $t-2$ to integrate out the unobserved regimes as follows

$$h_{t-1} = E_{t-2}\{h_{t-1}^{(j)}\} = p_{1,t-1}\left[\left(\mu_{t-1}^{(1)}\right)^2 + h_{t-1}^{(1)}\right] + (1-p_{1,t-1})\left[\left(\mu_{t-1}^{(2)}\right)^2 + h_{t-1}^{(2)}\right] - \left[p_{1,t-1}\mu_{t-1}^{(1)} + (1-p_{1,t-1})\,\mu_{t-1}^{(2)}\right]^2 \tag{3.7}$$

where $j = 1, 2$. The main drawback of this specification is its inconvenience in terms of volatility forecasting, since multi-step-ahead volatility forecasts turn out to be rather complicated. Dueker (1997) uses a collapsing procedure in the spirit of Kim's (1994) algorithm to overcome the path-dependence problem, but he essentially adopts the same framework of Gray (1996).

All these models have been put into a unified framework by Lin (1998) who gives the following specification for the conditional standard deviation σ_t

$$\frac{\sigma_t^{\nu}-1}{\nu} = \omega_{s_{t_1}} + \alpha_{s_{t_2}}(L)_p\,\widetilde{\sigma}_{t-1}^{\nu}\,|f(\varepsilon_{t-1})|^w - \lambda_{s_{t_2}}\widetilde{\sigma}_{t-1}^{\nu}\,|f(\varepsilon_{t-1})|^w\,\frac{\varepsilon_{t-1}}{|\varepsilon_{t-1}|} + \beta_{s_{t_3}}(L)_q\left[\frac{\widetilde{\sigma}_{t-1}^{\nu}-1}{\nu}\right] \tag{3.8}$$

where $t_1, t_2, t_3 \leq t$, $\widetilde{\sigma}_t$ denotes the conditional expectation of σ_t, $\alpha_{s_{t_2}}(L)_p$ and $\beta_{s_{t_3}}(L)_q$ represent polynomials in the lag operator (L) of order p and q respectively, and $f(\varepsilon_t) = \varepsilon_t - \gamma$. Lin (1998) also follows Gray's (1996) approach to avoid path-dependence.

Recently, Klaassen (2002) suggests to use the conditional expectation of the lagged conditional variance with a broader information set than in Gray (1996). To integrate out the past regimes taking into account also the current one, Klaassen (2002) adopts the following expression for the conditional variance

$$h_t^{(i)} = \alpha_0^{(i)} + \alpha_1^{(i)}\varepsilon_{t-1}^2 + \beta_1^{(i)}E_{t-1}\{h_{t-1}^{(i)}|s_t\} \tag{3.9}$$

where the expectation is computed as

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

$$E_{t-1}\{h_{t-1}^{(i)}|s_t\} = \sum_{j=1}^{2} \widetilde{p}_{ji,t-1} \left[\left(\mu_{t-1}^{(j)} \right)^2 + h_{t-1}^{(j)} \right] - \left[\sum_{j=1}^{2} \widetilde{p}_{ji,t-1} \mu_{t-1}^{(j)} \right]^2 \tag{3.10}$$

and the probabilities are calculated as

$$\widetilde{p}_{ji,t} = \Pr\left(s_t = j | s_{t+1} = i, \zeta_{t-1}\right) = \frac{p_{ji} \Pr\left(s_t = j|\zeta_{t-1}\right)}{\Pr\left(s_{t+1} = i|\zeta_{t-1}\right)} = \frac{p_{ji} p_{j,t}}{p_{i,t+1}} \tag{3.11}$$

with $i, j = 1, 2$.

Klaassen's (2002) regime-switching GARCH has two main advantages over the other MRS-GARCH models. Within the model, it allows higher flexibility in capturing the persistence of shocks to volatility.[5] Furthermore, it allows to have straightforward expressions for the multi-step-ahead volatility forecasts that can be calculated recursively as in standard GARCH models.

Since there is no serial correlation in the returns, the m-step-ahead volatility forecast at time $T-1$ can be calculated as follows

$$\widehat{h}_{T,T+m} = \sum_{\tau=1}^{m} \widehat{h}_{T,T+\tau} = \sum_{\tau=1}^{m} \sum_{i=1}^{2} \Pr\left(s_\tau = i|\zeta_{T-1}\right) \widehat{h}_{T,T+\tau}^{(i)} \tag{3.12}$$

where $\widehat{h}_{T,T+m}$ denotes the time T aggregated volatility forecast for the next m steps, and $\widehat{h}_{T,T+\tau}^{(i)}$ indicates the τ-step-ahead volatility forecast in regime i made at time T that can be calculated recursively

$$\widehat{h}_{T,T+\tau}^{(i)} = \alpha_0^{(i)} + \left(\alpha_1^{(i)} + \beta_1^{(i)}\right) E_T\{h_{T,T+\tau-1}^{(i)}|s_{T+\tau}\} \tag{3.13}$$

Therefore, the multi-step-ahead volatility forecasts are computed as a weighted average of the multi-step-ahead volatility forecasts in each regime, where the weights are the prediction probabilities. In each regime, the volatility forecast is obtained with the GARCH-like formula in (3.13), where the expectation of the previous period volatility is determined by weighting the precedent regime volatilities with the probabilities in (3.11). In general, to compute the volatility forecasts,

[5] A shock can be followed by a volatile period not only because of GARCH effects but also because of a switch to the higher variance regime. Having different parameters across regimes can capture the 'pressure-relieving' effect of some large shocks.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

the filter probability at τ periods ahead $\Pr(s_{t+\tau} = i|\zeta_t) = p_{i,t+\tau} = P^\tau p_{i,t}$ is needed.

Typically, in the Markov regime-switching literature maximum likelihood estimation (MLE) is adopted to estimate the numerous parameters. An essential ingredient is the ex-ante probability $p_{1,t} = \Pr[S_t = 1|\zeta_{t-1}]$, i.e. the probability of being in the first regime at time t given the information at $t-1$, whose specification is

$$p_{j,t} = \Pr[s_t = j|\zeta_{t-1}] = \sum_{i=1}^{2} p_{ij} \left[\frac{f(r_{t-1}|s_{t-1} = i)\, p_{i,t-1}}{\sum_{k=1}^{2} f(r_{t-1}|s_{t-1} = k)\, p_{k,t-1}} \right] \tag{3.14}$$

where p_{ij} are the transition probabilities in (3.2) and $f(\cdot)$ is the likelihood given in (3.3).

Thus, the log-likelihood function can be written as

$$\ell = \sum_{t=-R+w+1}^{T+w} \log[p_{1,t} f(r_t|s_t = 1) + (1 - p_{1,t}) f(r_t|s_t = 2)] \tag{3.15}$$

where $w = 0, 1, \ldots, n$, and $f(\cdot|s_t = i)$ is the conditional distribution given that regime i occurs at time t.

4 Data and Methodology

The data set analyzed in this paper is the Standard & Poor 100 (S&P100) stock market daily closing price index. The sample period is from January 1, 1988 to September 15, 2003 for a total of 4095 observations all obtained from Datastream. The sample is divided in two parts. The first 3585 observations (from January 1, 1988 to September 28, 2001) are used as the in-sample for estimation purposes, while the remaining 511 observations (from October 1, 2001 to September 15, 2003) are taken as the out-of-sample for forecast evaluation purposes.

Table 1 shows some descriptive statistics of the S&P100 rates of return. The mean is quite small (about 0.5%) and the standard deviation is around unity. The kurtosis is significantly higher than the normal value of 3 indicating that fat-tailed distributions are necessary to correctly describe r_t's conditional distribution. The skewness is significant, small and negative, showing that the lower tail of the

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 1: Descriptive Statistics

Mean	Standard Deviation	Min	Max	Skewness	Kurtosis	Normality Test	$LM(12)$	$Q^2(12)$
0.0359	1.0887	-7.6445	5.6901	-0.1972	7.3103	3214.44*	404.99*	863.69*

Note: The sample period is January 1, 1988 through September 15, 2003. The Normality Test is the Jarque-Bera test which has a χ^2 distribution with 2 degrees of freedom under the null hypothesis of normally distributed errors. The 5% critical value is, therefore, 5.99. The $LM(12)$ statistic is the ARCH LM test up to the twelfth lag and under the null hypothesis of no ARCH effects it has a $\chi^2(q)$ distribution, where q is the number of lags. The $Q^2(12)$ statistic is the Ljung-Box test on the squared residuals of the conditional mean regression up to the twelfth order. Under the null hypothesis of no serial correlation, the test is also distributed as a $\chi^2(q)$, where q is the number of lags. Thus, for both tests the 5% critical value is 21.03. At a confidence level of 5% both skewness and kurtosis are significant, since the standard errors under the null of normality are $\sqrt{6/T} = 0.038$ and $\sqrt{24/T} = 0.076$ respectively.

empirical distribution of the returns is longer than the upper tail, i.e. negative returns are more likely to be far below the mean than their counterparts.

In Table 1, LM(12) is the Lagrange Multiplier test for ARCH effects in the OLS residuals from the regression of the returns on a constant, while $Q^2(12)$ is the corresponding Ljung-Box statistic on the squared standardized residuals. Both these statistics are highly significant suggesting the presence of ARCH effects in the S&P100 returns up to the twelfth order.

The group of competing GARCH models with or without state-dependent parameters are estimated using quasi-maximum likelihood (QML). Both the conditional mean and the conditional variances are estimated jointly by maximizing the log-likelihood function computed as the logarithm of the product of the conditional densities of the prediction errors as in (3.15).

The ML estimates are obtained by maximizing the log-likelihood with the Broyden, Fletcher, Goldfarb, and Shanno (BFGS) quasi-Newton optimization algorithm in the MATLAB numerical optimization routines.[6]

The "true volatility" is needed to evaluate the forecasting performances of the competing GARCH models both in-sample and out-of-sample. So far in the literature, many researchers have used either the ex-ante or the ex-post squared returns in order to proxy the realized volatility. However, the squared returns represent

[6]Some of the iterative procedures have been written in C/C++ in order to enhance speed and to improve capabilities not directly available in MATLAB.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

a very noisy estimate of the unobserved volatility and can lead to wrong assessments about the real ability of various GARCH models to forecast volatility. As highlighted in Andersen and Bollerslev (1998) , a way to avoid such conclusions is using a more precise measure of volatility, obtained with intra-daily data. This measure is called 'realized volatility' and is based on the cumulative squared intra-daily returns over different time intervals from one minute to one hour.

In this paper we adopt three different measures of the actual volatility at time t, denoted $\widehat{\sigma}_t^2$. The first one is the realized volatility computed as the sum of one-minute squared returns over each day. Intra-daily returns on the S&P100 index are obtained from www.disktrading.com (these data are also used by Koopman *et al.*, 2005). To calculate the m-step-ahead volatility, we sum the daily realized volatilities over m days. The second measure is the more classical squared return for the daily volatility, which is summed over the relevant days for horizons greater than one day. The third measure is given by the squared return of the forecast horizon. Thus, if for example we are forecasting volatility at one week, we use the square of the log difference of the closing prices at time t and $t+5$.

We denote the m-step-ahead volatility forecast as $\widehat{h}_{t,t+m}$, computed as the aggregated sum of the forecasts for the next m steps made at time t, i.e. $\widehat{h}_{t,t+m} = \sum_{j=1}^{m} \widehat{h}_{t+j}$, where $\widehat{h}_{t+j}$ is the volatility forecast at day $t+j$. We thus compute the volatility forecasts at the one-, five-, ten- and twenty-two-day horizons by aggregating the volatility forecasts over the next 1, 5, 10 and 22 days. Actually, practitioners and risk managers are not particularly interested in the multi-step-ahead one-day volatility forecasts, such as the volatility $\widehat{h}_{t+22}$ at time $t+22$ made at t.

5 Evaluation of Volatility Forecasts

5.1 Standard Statistical Loss Functions

Forecast evaluation is a key step in any forecasting exercise. A popular metric to evaluate different forecast models is the minimization of a particular statistical loss function. However, the evaluation of the quality of competing volatility models can be very difficult because, as remarked by both Bollerslev *et al.* (1994) and Lopez (2001), there does not exist a unique criterion capable of selecting the best model. Many researchers have highlighted the importance of evaluating volatility forecasts by means of the real loss function faced by the final user. For example, Engle *et al.* (1993) and West *et al.* (1993) suggest profit-based and utility-based

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

criteria for evaluating the accuracy of volatility forecasts. Unfortunately, it is not possible to exactly know such loss function, because it depends on the unknown and unobservable economic agents' preferences. Thus, even though rather criticizable, so far most of the literature has focused on a particular set of statistical loss functions of the MSE-type.

In the present work, instead of choosing a particular statistical loss function as the best and unique criterion, we adopt seven different loss functions, that can have different interpretations and can lead to a more complete forecast evaluation of the competing models. These statistical loss functions are:

$$MSE_1 = n^{-1} \sum_{t=1}^{n} \left(\widehat{\sigma}_{t+m} - \widehat{h}_{t,t+m}^{1/2} \right)^2 \tag{5.1}$$

$$MSE_2 = n^{-1} \sum_{t=1}^{n} \left(\widehat{\sigma}_{t+m}^2 - \widehat{h}_{t,t+m} \right)^2 \tag{5.2}$$

$$QLIKE = n^{-1} \sum_{t=1}^{n} \left(\log \widehat{h}_{t,t+m} + \widehat{\sigma}_{t+m}^2 \widehat{h}_{t,t+m}^{-1} \right) \tag{5.3}$$

$$R2LOG = n^{-1} \sum_{t=1}^{n} \left[\log \left(\widehat{\sigma}_{t+m}^2 \widehat{h}_{t,t+m}^{-1} \right) \right]^2 \tag{5.4}$$

$$MAD_1 = n^{-1} \sum_{t=1}^{n} \left| \widehat{\sigma}_{t+m} - \widehat{h}_{t,t+m}^{1/2} \right| \tag{5.5}$$

$$MAD_2 = n^{-1} \sum_{t=1}^{n} \left| \widehat{\sigma}_{t+m}^2 - \widehat{h}_{t,t+m} \right| \tag{5.6}$$

$$HMSE = n^{-1} \sum_{t=1}^{n} \left(\widehat{\sigma}_{t+m}^2 \widehat{h}_{t,t+m}^{-1} - 1 \right)^2 \tag{5.7}$$

The criteria in (5.1) and (5.2) are the typical mean squared error metrics. The criteria in (5.2) and (5.4) are exactly equivalent to using the R^2 metric in the Mincer-Zarnowitz regressions of $\widehat{\sigma}_{t+m}^2$ on a constant and $\widehat{h}_{t,t+m}$ and of $\log\left(\widehat{\sigma}_{t+m}^2\right)$ on a constant and $\log\left(\widehat{h}_{t,t+m}\right)$, respectively, provided that the forecasts are unbiased. Moreover, the $R2LOG$ loss function has the particular feature of penalizing volatility forecasts asymmetrically in low and high volatility periods, as pointed out by Pagan and Schwert (1990) who put forward (5.4), calling it logarithmic loss

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

function. The loss function in (5.3) corresponds to the loss implied by a gaussian likelihood and is suggested by Bollerslev *et al.* (1994). The Mean Absolute Deviation (MAD) criteria in (5.5) and (5.6) are useful because they are generally more robust to the possible presence of outliers than the MSE criteria, but they impose the same penalty on over- and under-predictions and are not invariant to scale transformations. Bollerslev and Ghysels (1996) propose the heteroscedasticity-adjusted MSE in (5.7).

When comparing different volatility forecasts it can also be useful to measure the number of times a given model correctly predicts the directions of change[7] of the actual volatility. Such directional accuracy of volatility forecasts can be of great importance because the direction of predicted volatility change can be used to construct particular trading strategies such as straddles (Engle *et al.*, 1993).

Some tests of directional predictive ability have been proposed in the literature. In the present paper we use the so-called Success Ratio (SR) and the Directional Accuracy (DA) test of Pesaran and Timmermann (1992) .

Let $\overline{\sigma}_{t+m}$ be the proxy for the actual volatility at $t+m$ after subtracting its non-zero mean and let $\overline{h}_{t,t+m}$ be the demeaned volatility forecast.[8] The SR is simply the fraction of the volatility forecasts that have the same direction of change as the corresponding realizations and is given by

$$SR = n^{-1} \sum_{j=0}^{n-1} \mathcal{I}_{\{\overline{\sigma}_{t+m+j} \overline{h}_{t+j,t+m+j} > 0\}} \tag{5.8}$$

where $m = 1, 2, \ldots, 22$ and $\mathcal{I}_{\{g>0\}}$ is the indicator function, i.e. $\mathcal{I}_{\{g>0\}} = 1$ if g is positive and zero otherwise. Thus the SR measures the number of times the volatility forecast correctly predicts the direction of the true volatility process.

The DA test is instead given by

$$DA = \frac{(SR - SRI)}{\sqrt{Var(SR) - Var(SRI)}} \tag{5.9}$$

where $SRI = P\widehat{P} + (1-P)(1-\widehat{P})$ and P represents the fraction of times that $\overline{\sigma}_{t+m+j} > 0$, while $\widehat{P}$ is the proportion of demeaned volatility forecasts that are

[7] We refer to the direction of change, since volatility is always positive.

[8] The author acknowledges that because of the Jensen's inequality, $E(X^2) \geq [E(X)]^2$, such a procedure can give results that might be partially misleading. However, both the averages of $\widehat{\sigma}_t^2$ and $\widehat{h}_t$ can be overestimated, and particularly the former. Therefore, the results for the sign tests should be only partially underestimated.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

positive. $Var(SR)$ and $Var(SRI)$ are the corresponding variances. The DA test is asymptotically distributed as a standard normal.

5.2 Tests of Equal and Superior Predictive Ability

Forecasts from competing models are usually compared either with a pairwise or a joint test. When one compares the predictive ability of pairs of competing models, the usual test employed is the Diebold-Mariano (DM) Test or one of its modifications. However, it is far more sensible to compare the predictive ability of competing forecasts altogether, because with a pairwise comparison we can only test two different models and decide which one is better. The tests that we can adopt in this case are the Reality Check (RC) of White (2000) or the Superior Predictive Ability (SPA) test of Hansen (2005).

Diebold and Mariano (1995) propose a test of EPA of two competing models. Such a test is based on the null hypothesis of no difference in the accuracy of the two competing forecasts.

Assuming that the parameters of the system are set a priori and do not require estimation, the DM test statistic is designed as follows: let $\{\widehat{r}_{i,t}\}_{t=1}^{n}$ and $\{\widehat{r}_{j,t}\}_{t=1}^{n}$ denote two sequences of forecasts of the series $\{r_t\}_{t=1}^{n}$ generated by two competing models i and j and let $\{e_{i,t}\}_{t=1}^{n}$ and $\{e_{j,t}\}_{t=1}^{n}$ be the corresponding forecast errors. Assuming that the loss function $g(\cdot)$ can be written as a function of only the forecast errors, we can define the loss differential between the two competing forecasts as $d_t \equiv [g(e_{i,t}) - g(e_{j,t})]$. Then, assuming that the sequence $\{d_t\}_{t=1}^{n}$ is covariance stationary and has a short memory, Diebold and Mariano (1995) show that the asymptotic distribution of the sample mean loss differential $\overline{d} = \frac{1}{n}\sum_{t=1}^{n} d_t$ is $\sqrt{n}\left(\overline{d} - \mu\right) \xrightarrow{d} N\left(0, V\left(\overline{d}\right)\right)$. An estimate of the asymptotic variance is $\widehat{V}\left(\overline{d}\right) = n^{-1}\left(\widehat{\gamma}_0 + 2\sum_{k=1}^{q}\omega_k\widehat{\gamma}_k\right)$, where $q = h-1$, $\omega_k = 1-k/(q+1)$ is the lag window and $\widehat{\gamma}_i$ is an estimate of the i-th order autocovariance of the series $\{d_t\}_1^n$ that can be estimated as $\widehat{\gamma}_k = \frac{1}{n}\sum_{t=k+1}^{n}\left(d_t - \overline{d}\right)\left(d_{t-k} - \overline{d}\right)$ for $k = 1, \ldots, q$.[9] The DM statistic to test the null hypothesis of equal forecast accuracy is then given by $DM = \overline{d}/\sqrt{\widehat{V}\left(\overline{d}\right)} \sim N(0,1)$, i.e. under the null hypothesis of equal forecast accuracy the DM test statistic has a standard normal distribution asymptotically. Harvey *et al.* (1997) argue that the DM test can be quite over-sized in small samples and this problem becomes even more dramatic as the forecast horizon increases. They thus suggest a Modified DM (MDM)

[9] q has been chosen so that $q = \lfloor 4 * (n/100)^{2/9} \rfloor$.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

test, where DM is multiplied by the factor $\sqrt{n^{-1}\left[n+1-2m+n^{-1}m(m-1)\right]}$, where m is the forecast horizon and n is the length of the evaluation period.[10]

Instead of testing for EPA, as in Diebold and Mariano (1995) and West (1996), the RC for data snooping of White (2000) is a test for superior predictive ability.[11]

The RC is constructed in a way to test whether a particular forecasting model is significantly outperformed by a set of alternative models, where the performance of each model may be defined according to a pre-specified loss function.

White (2000) compares $l+1$ forecasting models. Model 0 is the benchmark and the null hypothesis is that none of the models $k=1,\ldots,l$ outperforms the benchmark in terms of the specific loss function chosen. The best forecast model is that one which produces the smallest expected loss. Let $L_{t,k} \equiv L(\widehat{\sigma}_t^2, \widehat{h}_{k,t})$ denote the loss[12] if one makes the prediction $\widehat{h}_{t,k}$ with k-th model when the realized volatility turns out to be $\widehat{\sigma}_t^2$. The performance of model k relative to the benchmark model (at time t), can be defined as

$$f_{k,t} = L_{t,0} - L_{t,k} \quad k=1,\ldots,l; \quad t=1,\ldots,n \tag{5.10}$$

Assuming stationarity for $f_{k,t}$ we can define the expected relative performance of model k relative to the benchmark as $\mu_k = E[f_{k,t}]$ for $k=1,\ldots,l$. If model w outperforms the benchmark, then the value of μ_w will be positive. Therefore, we can analyze whether any of the competing models significantly outperform the benchmark, testing the null hypothesis that $\mu_k \leq 0$, for $k=1,\ldots,l$. Consequently, the null hypothesis that none of the models is better than the benchmark (i.e. no predictive superiority over the benchmark itself) can be equivalently formulated as

$$H_0 : \mu_{\max} \equiv \max_{k=1,\ldots,l} \mu_k \leq 0 \tag{5.11}$$

against the alternative that the best model is superior to the benchmark.

By the law of large numbers we can consistently estimate μ_k with the sample average $\overline{f}_{k,n} = n^{-1}\sum_{t=1}^{n} f_{k,t}$ and then obtain the test statistic

[10] Harvey *et al.* (1997) suggest to compare the statistic with the critical values from the Student's t distribution with $n-1$ degrees of freedom rather than from the normal distribution as with the DM test.

[11] In economics, testing for SPA is certainly more relevant than testing for EPA, because we are more interested in the possibility of the existence of the best forecasting model rather than in the probable existence of a better model between two pairs.

[12] The function $L(\cdot)$ can be anyone of the loss functions given before. For example, it can be $L_{k,t} = (\widehat{\sigma}_t^2 - \widehat{h}_{t,k})^2$ if we consider the loss function in (5.2).

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

$$T_n \equiv \max_{k=1,\dots,l} n^{1/2}\overline{f}_{k,n} \tag{5.12}$$

If we reject the null hypothesis, we have evidence that among the competing models, at least one is significantly better than the benchmark.

The most difficult problem is to derive the distribution of the statistic T_n under H_0, because the distribution is not unique. Hansen (2005) emphasizes that the RC test applies a supremum over the non-standardized performances T_n and, more dangerously, a conservative asymptotic distribution that makes it very sensitive to the inclusion of poor models. The author argues that since the distribution of the statistic is not unique under the null hypothesis, it is necessary to obtain a consistent estimate of the p-value, as well as a lower and an upper bound. Therefore, he applies a supremum over the standardized performances and tests the null hypothesis

$$H_0 : \ \mu^s_{\max} \equiv \max_{k=1,\dots,l} \frac{\mu_k}{\sqrt{var(n^{1/2}\overline{f}_{k,n})}} \leq 0 \tag{5.13}$$

using the statistic

$$T^s_n = \max_k \frac{n^{1/2}\overline{f}_{k,t}}{\sqrt{\widehat{var}(n^{1/2}\overline{f}_{k,n})}} \tag{5.14}$$

where $\widehat{var}(n^{1/2}\overline{f}_{k,n})$ is an estimate of the variance of $n^{1/2}\overline{f}_{k,n}$ obtained via the bootstrap. Therefore, Hansen (2005) suggests additional refinements to the RC test and some modifications of its asymptotic distribution that result in tests less sensitive to the inclusion of poor models and with a better power. He also argues that the p-values of the RC are generally inconsistent (i.e. too large) and the test can be asymptotically biased. To overcome these drawbacks, Hansen (2005) shows that it is possible to derive a consistent estimate of the p-value along with an upper and a lower bound. Such a test is called SPA test and it includes the RC as a special case. The upper bound (SPA_u) is the p-value of a conservative test (i.e. it has the same asymptotic distribution as the RC test) where it is implicitly assumed that all the competing models $(k = 1, \dots, l)$ are as good as the benchmark in terms of expected loss. Hence, the upper-bound p-value coincides with that one of the RC. The lower-bound (SPA_l) is the p-value of the liberal test where the null hypothesis assumes that those models with worse performance than the benchmark are poor models in the limit. With the SPA test it is possible

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

to assess which models are worse than the benchmark and asymptotically we can prevent them from affecting the distribution of the test statistic. The conservative test (and thus the RC test) is quite sensitive to the inclusion of poor and irrelevant models in the comparison, while the consistent (SPA_c) and the liberal tests are not.[13]

5.3 Risk-Management Loss Functions

Since volatility forecasts are typically used as an input to financial risk management, we also employ a risk-management loss function based upon the calculation of the VaR. An institution's VaR is a measure of the market risk of a portfolio which quantifies in monetary terms the likely losses which could arise from market fluctuations. Brooks and Persand (2003) and Sarma *et al.* (2003) suggest to use VaR-based loss functions to evaluate competing models.

In a regime-switching framework (see Billio and Pellizzon, 2000 or Guidolin and Timmermann, 2004) the VaR at time t of model k at $\alpha\%$ significance level is calculated as follows

$$VaR_t^k\left[m,\alpha\right]=\sum_{i=1}^{2}\Pr\left[s_{t+m}=i|\zeta_{t-1}\right]\left[\widehat{\mu}_{t+m,k}^{(i)}+\Phi\left(\alpha\right)\left(\widehat{h}_{t+m,k}^{(i)}\right)^{1/2}\right] \quad (5.15)$$

where $\Phi(\cdot)$ is a cumulative distribution function, m is the investment horizon ($m=1,5,10,22$ days), α =1% or 5%, $\widehat{\mu}_{t+m,k}^{(i)}$ is the mean forecast and $\widehat{h}_{t+m,k}^{(i)}$ is the aggregate volatility forecast at $t+m$ in regime i from the k-th model.

To compare the forecasting ability of different models in terms of VaR and thus determine the adequacy of their volatility forecasts as an input for the VaR, we can use a group of tests which is quite standard in the literature.

The first test is based upon the number of observations before the first exception, i.e. the Time Until First Failure ($TUFF$). First, we construct the failure process, which indicates the exact number of exceptions of the VaR from model k at time t, i.e. $I_t=\mathcal{I}_{\{r_t<VaR_t^k\}}$, where $\mathcal{I}_{\{\}}$ is the usual indicator function. The relevant null hypothesis is $H_0:\alpha=\alpha_0$ and the corresponding likelihood ratio (LR) test is

[13]For a detailed description of how to implement the RC and SPA test, see White (2000) and Hansen (2005).

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

$$LR_{TUFF}(\widetilde{T},\widehat{\alpha}) = -2\log\left[\widehat{\alpha}\left(1-\widehat{\alpha}\right)^{\widetilde{T}-1}\right] + 2\log\left[\widetilde{T}^{-1}\left(1-\widetilde{T}^{-1}\right)^{\widetilde{T}-1}\right] \quad (5.16)$$

where $\widetilde{T}$ denotes the number of observations before the first exception. LR_{TUFF} is asymptotically distributed as $\chi^2(1)$ under the null. Thus, the 95% confidence interval for $TUFF$ is (3,514) for the 99% VaR and (1,101) for the 95% VaR. The interpretation of such confidence intervals is straightforward. If the 99% VaR fails before the third observation, we can reject at 5% the null that the model is adequate to cover losses at 99%. Conversely, if the $TUFF$ is greater than 514, then we would conclude that the model is inadequate because it suggests a VaR which is too high. Either cases would be detrimental for an institution because with a too low VaR the capital could not be sufficient to cover future losses, while, on the other side, a too high VaR would imply that too much capital is unprofitably tied up. In addition, Kupiec (1995) notices that this test has limited power to distinguish among alternative hypotheses since all observations after the first failure are ignored, resulting in a test which is over-sized. Therefore, it would be unacceptable to rely only on this criterion for model adequacy because the test has low power to reject poor models.

In this risk-management framework we have another set of tests for model adequacy based on the percentage of times that the calculated VaR is insufficient to cover the actual losses. A model is judged as adequate if its proportion of failures (PF) in the out-of-sample is close to the nominal value (5% or 1%). This way of testing can be carried out either unconditionally or conditionally. Unconditionally, the VaR is said to be efficient, or equivalently to display correct unconditional coverage, if $E\left(I_t\right) = \alpha$. In other words, under the assumption that the number of failures in the hold-out sample is *iid* binomially distributed, a LR test of unconditional coverage is

$$LR_{UC} = LR_{PF} = -2\log\left[\frac{\alpha^{n_1}\left(1-\alpha\right)^{n_0}}{\widehat{\alpha}^{n_1}\left(1-\widehat{\alpha}\right)^{n_0}}\right] \sim \chi^2_{(1)} \quad (5.17)$$

where α is the tolerance level at which VaR measures are estimated (i.e. 1 or 5 %), n_1 is the number of 1's - i.e. failures - in I_t, n_0 is the number of 0's, and $\widehat{\alpha} = n_1/(n_0+n_1)$ is the MLE estimate of α.

However, Christoffersen (1998) suggests that a correctly specified VaR model should generate the pre-specified failure rate conditionally at every point in time. This property is called conditional coverage and finds its origin in the well known

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

stylized fact of volatility clustering. The author develops a framework for interval forecast evaluation arguing that good interval forecasts should be narrow in tranquil periods and wide in volatile times, and thus observations falling outside a forecasted interval should be evenly spread over the entire sample rather than clustered. The test of conditional coverage combines the test of unconditional coverage with a test of independence, where the null of an independently distributed failure process is tested against the alternative of a first order Markov failure process. The LR statistic of the test of independence is

$$LR_{IND} = -2\log\left[\frac{(1-\widehat{\pi})^{(n_{00}+n_{10})}(1-\widehat{\pi})^{(n_{01}+n_{11})}}{(1-\widehat{\pi}_{01})^{n_{00}}\widehat{\pi}_{01}^{n_{01}}(1-\widehat{\pi}_{11})^{n_{10}}\widehat{\pi}_{11}^{n_{11}}}\right] \sim \chi^2_{(1)} \tag{5.18}$$

where n_{ij} is the number of i values followed by a j value in the failure process I_t with $i,j=0,1$, $\pi_{ij}=\Pr\{I_t=i|I_{t-1}=j\}$ with $i,j=0,1$, $\widehat{\pi}_{01}=n_{01}/(n_{00}+n_{01})$, $\widehat{\pi}_{11}=n_{11}/(n_{10}+n_{11})$, $\widehat{\pi}=(n_{01}+n_{11})/(n_{00}+n_{01}+n_{10}+n_{11})$.

Finally, the test of correct conditional coverage is performed by testing the null of an independent failure process with failure probability α against the alternative of a first order Markov failure process. The LR statistic for the test is

$$LR_{CC} = -2\log\left[\frac{(1-\alpha)^{n_0}\alpha^{n_1}}{(1-\widehat{\pi}_{01})^{n_{00}}\widehat{\pi}_{01}^{n_{01}}(1-\widehat{\pi}_{11})^{n_{10}}\widehat{\pi}_{11}^{n_{11}}}\right] \sim \chi^2_{(2)} \tag{5.19}$$

If we condition on the first observation, then these LR test statistics are related by the identity $LR_{CC}=LR_{UC}+LR_{IND}$.

For those models which can pass these tests of correct coverage we follow Sarma *et al.* (2003) by further evaluating the competing models through VaR-based loss functions which are closer to the real risk managers' utility functions. These authors suggest using Lopez's (1999) approach by incorporating the magnitude of the exceptions in addition to their number and by penalizing failures asymmetrically. They propose a regulator loss function (RLF), given by $l_t^1 = \left(r_t - VaR_t^k\right)^2 \mathcal{I}_{\{r_t<VaR_t^k\}}$, and a firm loss function (FLF), given by $l_t^2 = \left(r_t - VaR_t^k\right)^2 \mathcal{I}_{\{r_t<VaR_t^k\}} - \delta VaR_t^k \mathcal{I}_{\{r_t>VaR_t^k\}}$, where $\mathcal{I}$ is the usual indicator function and δ is the opportunity cost of capital.[14] To assess the superiority of one model to another, Sarma *et al.* (2003) adopt the same one-sided sign test as

[14]In what follows, we assume that the opportunity cost of capital can be linked to the risk-free interest rate and thus we set $\delta = 1.5\%$.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

in Diebold and Mariano (1995). Defining the loss differential between model i and j as $z_t = l_{it}^k - l_{jt}^k$, they test the null of a zero-median loss differential against the alternative of a negative median, with a studentized version of the sign test given by $\widehat{S}_{ij} = (S_{ij} - 0.5T)(0.25T)^{-.5}$, where $S_{ij} = \sum_{t=1}^{T} \mathcal{I}_{\{z_t>0\}}$. Under the null, $\widehat{S}_{ij}$ is asymptotically distributed as a standard normal. Thus, if $\widehat{S}_{ij} < -1.645$ one could then reject the null of a nil median against a negative one, which would imply that model i is significantly better than model j.

In the VaR-based forecast evaluation we use all these tests so that competing models are evaluated as follows. First, we comment those models that present a $TUFF$ which is either too low or too high with respect to the 95% confidence interval for LR_{TUFF}, but we do not reject them as inadequate because this test has low power to reject models that are inadequate.[15] Then, as in Brooks and Persand (2003) we assume that any model which has a percentage of failures in the rolling hold-out sample which is either far greater than the nominal threshold $\alpha\%$ or far less should be rejected as inadequate. We thus select those models with a PF which is closer to the nominal level, i.e. those models which have the lowest values of LR_{PF}. Afterwards, we reject as inadequate all models that lack the property of correct conditional coverage (i.e. that have significant LR_{CC}). Finally, for those models that have passed all these coverage tests, we compute the nonparametric sign test discussed above to check whether the losses from two competing models are significantly different from each other, thus selecting the model that is superior in each forecast horizon. During this final step, we compare different pairs of best models under both RLF and FLF. First, we compare those models that exhibit the lowest average RLF and FLF within the group of standard GARCH and MRS-GARCH models. Then, we compare all the pairs of models within each group that show the lowest values[16] for both LR_{PF} and LR_{CC}. If a model turns out to be significantly different from the others, it is considered adequate for risk-management purposes.

[15] In addition, our out-of-sample period starts out in a period of exceptional market turbulence, that might lead to too early failures for most models.

[16] In case we reject either conditional or unconditional coverage for all models, we should stop the evaluation. Nevertheless, in the empirical example below, we continue the comparison under both RLF and FLF for the sake of completeness.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

6 Empirical Results and Discussion

The whole sample consists of the S&P100 closing prices from January 1, 1988 to September 15, 2003, for a total of 4095 observations. The return series is calculated by taking the log difference of price indices and then multiplying by 100. The estimation is carried out on a moving (or rolling) window of 3585 observations. In this section we present the empirical estimates of single-regime GARCH and MRS-GARCH models, along with the in-sample statistics and the out-of-sample forecast evaluation.

6.1 Single-regime GARCH

The parameter estimates of the different state-independent GARCH(1,1) models are presented in Table 2. For each model three different distributions for the innovations are considered: the Normal, the Student's t and the GED. The in-sample period is from January 1, 1988 through September 28, 2001. The 511 observations from October 1, 2001 through September 15, 2003 are reserved for the evaluation of the out-of-sample performances. The standard errors are the asymptotic standard errors. Regarding the conditional mean, all the parameters of the various GARCH models are significant. The conditional variance estimates show that almost all the parameters are highly significant, except for the α_0's in the GARCH and GJR models. Hence GARCH models perform quite well at least in-sample. In addition, for the Student's t distribution, the degrees of freedom are always greater than 6, suggesting that all the conditional moments up to the sixth order exist. In particular, the conditional kurtosis of the Student's t distribution is given by $3(\nu-2)/(\nu-4)$. Consequently, in the GARCH, EGARCH and GJR model, the value for the conditional kurtosis is 5.538, 5.003 and 5.009 respectively, confirming the typical fat-tailed behavior of financial returns.

Moreover, for the models with GED innovations, the estimates clearly suggest that the conditional distribution has fatter tails than the gaussian, since all the shape parameters have values that significantly lie between 1 and 2. The same conclusion arises with the conditional kurtosis given by $\left(\Gamma\left(1/\nu\right)\Gamma\left(5/\nu\right)\right)/\left(\Gamma\left(1/\nu\right)\right)^2\right)$, where $\Gamma(\cdot)$ is the gamma function. For the GARCH, EGARCH and GJR model the kurtosis is 4.149, 4.026 and 4.044 respectively, confirming that the estimated conditional distribution of S&P100 returns is indeed fat-tailed.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 2: Maximum Likelihood Estimates of Standard GARCH Models.

	GARCH			EGARCH			GJR		
	N	t	GED	N	t	GED	N	t	GED
δ	0.0562	0.0610	0.0441	0.0362	0.0453	0.0305	0.0382	0.0500	0.0340
	(0.0140)	(0.0130)	(0.0120)	(0.0140)	(0.0130)	(0.0120)	(0.0150)	(0.0130)	(0.0120)
α_0	0.0220	0.0182	0.0187	-0.0747	-0.0775	-0.0745	0.0285	0.0209	0.0230
	(0.0020)	(0.0030)	(0.0030)	(0.0050)	(0.0100)	(0.0100)	(0.0020)	(0.0030)	(0.0040)
α_1	0.0752	0.0751	0.0746	0.0975	0.1021	0.0985	0.1293	0.1289	0.1300
	(0.0050)	(0.0100)	(0.0100)	(0.0070)	(0.0140)	(0.0140)	(0.0090)	(0.0150)	(0.0150)
β_1	0.9017	0.9049	0.9046	0.9855	0.9900	0.9889	0.8977	0.9022	0.9011
	(0.0060)	(0.0090)	(0.0100)	(0.0020)	(0.0100)	(0.0030)	(0.0070)	(0.0100)	(0.0110)
ξ	-	-	-	-0.0598	-0.0635	-0.0614	0.0141	0.0203	0.0186
				(0.0050)	(0.0030)	(0.0100)	(0.0070)	(0.0110)	(0.0120)
ν	-	5.4416	1.2047	-	5.4469	1.2162	-	5.7332	1.2259
		(0.4640)	(0.0290)		(0.4680)	(0.0280)		(0.5040)	(0.0290)
$Log(L)$	-4816.3791	-4671.9133	-4668.2808	-4777.9987	-4630.1921	-4633.4375	-4780.2798	-4649.2711	-4646.1104

Note: Each GARCH model has been estimated with a Normal (N), a Student's t and a GED distribution. The in-sample data consist of S&P100 returns from 1/1/1988 to 9/28/2001. The conditional mean is $r_t = \delta + \varepsilon_t$. The conditional variances are $h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta h_{t-1}$, $\log(h_t) = \alpha_0 + \alpha_1 \left| \frac{\varepsilon_{t-1}}{h_{t-1}} \right| + \xi \frac{\varepsilon_{t-1}}{h_{t-1}} + \beta_1 \log(h_{t-1})$, and $h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \left[1 - \mathcal{I}_{\{\varepsilon_{t-1}>0\}}\right] + \xi \varepsilon_{t-1}^2 \mathcal{I}_{\{\varepsilon_{t-1}>0\}} + \beta_1 h_{t-1}$ for the GARCH, EGARCH and GJR, respectively. Asymptotic standard errors are in parentheses.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

6.2 MRS-GARCH

The parameter estimates of MRS-GARCH models are presented in Table 3. Both the models with constant degrees of freedom and the one where the degrees-of-freedom parameters are allowed to switch between the two regimes show highly significant in-sample estimates. The conditional mean estimates are all significant, whereas for almost half of the conditional variance parameters, especially the constant $\alpha_0^{(i)}$'s, we fail to reject the null of a zero value. In particular, in Table 3 we report the standard deviation of the returns conditional on each volatility regimes, i.e. $\sigma^{(i)} = \left(\alpha_0^{(i)} / \left(1 - \alpha_1^{(i)} - \beta_1^{(i)}\right)\right)^{1/2}$, which is easier to interpret. The estimates confirm the existence of two states: the first regime is characterized by a low volatility and in most cases by a lower persistence of the shocks as indicated by $\rho_i = \alpha_1^{(i)} + \beta_1^{(i)}$. On the other hand, the second regime reveals a higher volatility and, almost always, a higher persistence. However, the overall persistence is never below .89. The transition probabilities are all highly significant and close to one except for the normal case where one of them is rather far away from unity, showing that almost all regimes are particularly persistent.

Table 3 also reports the unconditional probabilities of each MRS-GARCH model. The unconditional probability π_1 of being in the first regime, which is characterized by a lower volatility than the second, ranges between 33% for the model with GED innovations and 98% for the model with gaussian innovations. On the other hand, the unconditional probability of being in the high-volatility regime (the second one) ranges between 2% for the model with Normal innovations and 67% for the one with GED innovations. For the Student's t version of the MRS-GARCH with constant degrees of freedom across regimes, the shape parameter is below four, indicating the existence of conditional moments up to the third. This means that by allowing state-dependent parameters it is possible to model most of the leptokurtosis in the data. In the GED case the parameter ν is below the threshold value of 2, showing that the distribution has thicker tails than the normal. The conditional kurtosis in the GED case is 5.134.

The MRS-GARCH model with Student's t innovations is presented in the version in which the degrees of freedom switch across regimes, implying a time-varying kurtosis as in Hansen (1994) and Dueker (1997). However, there is a main difference between those papers and the present work. While Hansen suggests a model in which the degrees-of-freedom parameter varies over time according to a logistic function of variables included in the information set up to time $t-1$, and Dueker allows only such a parameter to be state-dependent, in the present paper

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 3: Maximum Likelihood Estimates of MRS-GARCH Models.

	MRS-GARCH-N	MRS-GARCH-t2	MRS-GARCH-t	MRS-GARCH-GED
$\delta^{(1)}$	0.0592	0.0571	0.0487	0.0252
	(0.0140)	(0.0140)	(0.0130)	(0.0120)
$\delta^{(2)}$	-1.6623	0.0558	0.0792	0.0715
	(0.2090)	(0.0310)	(0.0290)	(0.0210)
$\sigma^{(1)}$	0.3026	0.6672	0.5681	0.5621
	(0.2003)	(0.2668)	(0.1678)	(0.1941)
$\sigma^{(2)}$	6.4861	1.2930	1.4868	1.3751
	(3.4779)	(0.4160)	(0.3312)	(0.3149)
$\alpha_1^{(1)}$	0.0230	0.0137	0.0413	0.0388
	(0.0070)	(0.0050)	(0.0120)	(0.0150)
$\alpha_1^{(2)}$	0.0225	0.0754	0.0565	0.0560
	(0.1170)	(0.0190)	(0.0160)	(0.0140)
$\beta_1^{(1)}$	0.9093	0.9805	0.8473	0.8533
	(0.0090)	(0.0060)	(0.0330)	(0.0370)
$\beta_1^{(2)}$	0.9633	0.8569	0.8924	0.8952
	(0.1810)	(0.0290)	(0.0210)	(0.0200)
p	0.9811	0.9987	0.9998	0.9998
	(0.0040)	(0.0010)	(0.0001)	(0.0002)
q	0.1533	0.9988	0.9999	0.9999
	(0.0850)	(0.0010)	(0.0001)	(0.0001)
$\nu^{(1)}$	-	4.7318	5.3826	1.2212
		(0.5370)	(0.4440)	(0.0350)
$\nu^{(2)}$	-	7.1652	-	
		(1.3270)		
$Log(L)$	-4698.6929	-4632.0178	-4631.4338	-4630.9573
N. of Par.	10	12	11	11
π_1	0.98	0.48	0.39	0.33
π_2	0.02	0.52	0.61	0.67
ρ_1	0.93	0.99	0.89	0.89
ρ_2	0.99	0.93	0.95	0.95

Note: Each MRS-GARCH model has been estimated with different conditional distributions (see Section 3). The in-sample data consist of S&P100 returns from 1/1/1988 to 9/28/2001. The superscripts indicate the regime. The conditional mean is $r_t = \delta^{(i)} + \varepsilon_t$, whereas the conditional variance is $h_t^{(i)} = \alpha_0^{(i)} + \alpha_1^{(i)}\varepsilon_{t-1}^2 + \beta_1^{(i)} E_{t-1}\{h_{t-1}^{(i)}|s_t\}$, where the expectation is calculated as in (3.10). Instead of $\alpha_0^{(i)}$, we report $\sigma^{(i)} = \left(\alpha_0^{(i)} / \left(1 - \alpha_1^{(i)} - \beta_1^{(i)}\right)\right)^{1/2}$ for each regime which is the standard deviation conditional to the volatility regime. π_j is the unconditional probability of being in regime j, while $\rho_i = \alpha_1^{(i)} + \beta_1^{(i)}$ is the persistence of shocks in the i-th regime. Asymptotic standard errors are in parentheses.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

this parameter switches across regimes along with all the remaining parameters. Since both regimes show an estimated number of degrees of freedom greater than 4, we can argue that in both we have fatter tails than the normal.

6.3 In-Sample statistics

A big problem arises when one attempts to compare single-regime with regime-switching GARCH models. Standard econometric tests for model specification may not be appropriate because some parameters are unidentified under the null.[17] Since the main focus is on predictive ability, we only present some statistics in Table 4, without doing any formal test.

In Table 4 some in-sample goodness-of-fit statistics are reported. These statistics are used as model selection criteria. The largest log-likelihood among the state-independent GARCH models is given by the EGARCH model with GED innovations, while for the MRS-GARCH models, and overall, the best result is reached with the MRS-GARCH with Student's t distribution, where the degrees of freedom switch across the two volatility regimes.

The Akaike Information Criterion (AIC) and the Schwarz Criterion (BIC) both indicate that the best model among the standard GARCH and overall is the EGARCH with GED errors, while among the MRS-GARCH models is the MRS-GARCH-t2 that fits the best. Another property of MRS-GARCH models that emerges from Table 4 is the high persistence of the shocks to the conditional variance which is not so tiny as expected. Only in one regime the persistence is slightly smaller than in standard GARCH models. Table 4 also shows that according to all the statistical loss functions considered but for HMSE the best model in-sample is the MRS-GARCH with gaussian innovations, while among the standard GARCH models the best one is the EGARCH with normal innovations.

6.4 Out-of-Sample forecast evaluation

One possible way to overcome the problems highlighted in the previous section is to compare the models through their out-of-sample forecasting performances. An out-of-sample test has the ability to control either possible over-fitting or over-parametrization problems, and gives a more powerful framework to evaluate the performances of the competing models.

[17] See Hansen (1992 and 1996) who proposes simulation-based tests that can avoid this problem.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 4: In-sample goodness-of-fit statistics.

Model	N. of Par.	Pers.	AIC	Rank	BIC	Rank	$Log(L)$	Rank	MSE_1	Rank	MSE_2	Rank	$QLIKE$	Rank	$R2LOG$	Rank	MAD_2	Rank	MAD_1	Rank	$HMSE$	Rank
GARCH-N	4	0.977	2.689	13	2.696	13	-4816.379	13	0.599	11	9.291	10	0.848	8	8.635	11	1.227	11	0.587	10	6.382	3
GARCH-t	5	0.98	2.609	7	2.618	6	-4671.913	7	0.612	13	10.205	12	0.849	9	8.569	9	1.245	13	0.586	9	6.775	6
GARCH-GED	5	0.979	2.607	5	2.616	5	-4668.281	6	0.608	12	9.939	11	0.85	10	8.568	8	1.239	12	0.585	8	6.77	5
EGARCH-N	5	0.986	2.668	11	2.677	11	-4777.999	11	0.525	2	7.389	3	0.828	2	8.455	3	1.117	2	0.565	2	7.059	8
EGARCH-t	6	0.894	2.652	10	2.663	10	-4748.077	10	0.557	4	7.597	6	0.904	13	9.164	13	1.15	4	0.598	13	5.502	1
EGARCH-GED	6	0.989	2.588	1	2.599	1	-4633.437	2	0.53	3	7.386	2	0.829	3	8.432	2	1.125	3	0.566	3	7.365	10
GJR-N	5	0.969	2.67	12	2.678	12	-4780.28	12	0.57	5	8.22	7	0.829	4	8.561	7	1.192	5	0.578	4	6.439	4
GJR-t	6	0.977	2.597	4	2.607	3	-4649.271	4	0.593	10	9.279	9	0.83	6	8.493	4	1.225	10	0.58	7	7.157	9
GJR-GED	6	0.975	2.595	3	2.606	2	-4646.11	3	0.586	9	9.002	8	0.83	5	8.498	5	1.215	9	0.579	5	7.035	7
MRS-GARCH-N	10	0.986	2.627	9	2.644	9	-4698.693	9	0.524	1	7.33	1	0.846	7	8.312	1	1.113	1	0.559	1	8.83	12
MRS-GARCH-t2	12	0.994	2.591	2	2.612	4	-4632.018	1	0.579	6	10.697	13	0.821	1	8.554	6	1.204	6	0.58	6	5.782	2
MRS-GARCH-t	11	0.951	2.608	6	2.627	7	-4663.811	5	0.583	8	7.52	5	0.86	11	8.691	12	1.209	7	0.596	12	7.988	11
MRS-GARCH-GED	11	0.949	2.613	8	2.632	8	-4673.056	8	0.582	7	7.498	4	0.867	12	8.576	10	1.211	8	0.593	11	9.559	13

Note: Pers. is the persistence of shocks to volatility (for MRS-GARCH only the highest persistence is reported). AIC is the Akaike information criterion calculated as $-2\log(L)/T + 2k/T$, where k is the number of parameters and T the number of observations. BIC is the Schwarz criterion, calculated as $-2\log(L)/T + (k/T)\log(T)$. MSE_1, MSE_2, $QLIKE$, $R2LOG$, MAD_1, MAD_2, and $HMSE$ are the statistical loss functions introduced in Section 5.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Since most models only represent simple approximations of the true data generating process, often having a good in-sample fit does imply neither a necessary nor a sufficient condition for accurate and reliable forecasts. Furthermore, researchers and practitioners are particularly interested in having good volatility forecasts rather than good in-sample fits that might be much more likely with highly parameterized models such as MRG-GARCH.

Table 5 reports the out-of-sample evaluation of the one- and five-step-ahead volatility forecasts, according to the statistical loss functions in Section 5. Table 6 displays the out-of-sample evaluation of the ten- and twenty-two-step-ahead volatility forecasts. For both tables, the volatility proxy is given by the realized volatility.

All models exhibit a high SR (more than 60% and an average of 80%) and highly significant DA test at all forecast horizons.

At one day, the best model is the MRS-GARCH-N and the second best model is the GJR-N. At one week, the best model is again the MRS-GARCH-N, while the second best is the EGARCH-N. At the two-week horizon, the best model is the EGARCH-N and the MRS-GARCH-N is just the best model among the MRS-GARCH but it ranks the third. At the one-month horizon, the best model is the EGARCH-GED, while the MRS-GARCH-N is only the best among the MRS-GARCH ranking the sixth.[18]

From the previous results, it is quite evident that MRS-GARCH fare better at shorter forecast horizons, while at longer ones (more than a week) EGARCH and GJR models with non-normal innovations are the best. This is confirmed by the DM test for EPA of which, for the sake of brevity, we only present the tables when the benchmark is the MRS-GARCH-N at the one-day horizon and the EGARCH-N at the two-week horizon.

[18] When the proxy for the volatility is the d-day squared return, at all forecast horizons the best model is the EGARCH-t while the second best is the MRS-GARCH-N. When the volatility proxy is given by the sum of the daily squared returns, at the one-day horizon the best model is the GJR-N and the best among the MRS-GARCH (MRS-GARCH-N) is just the sixth. At the one-week horizon, the best model is the GJR-t, whereas the best among the MRS-GARCH (MRS-GARCH-t2) is the eighth. At the two-week horizon, the best model is the GJR-t, while the best among the MRS-GARCH (MRS-GARCH-GED) is the sixth. At the one-month horizon, the best model is the GJR-t and the best among the MRS-GARCH (MRS-GARCH-GED) is the fourth. All the corresponding tables are available upon request from the author.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 5: Out-of-sample evaluation of the one- and five-step-ahead volatility forecasts.

1-step-ahead volatility forecasts

Model	MSE_1	Rank	MSE_2	Rank	$QLIKE$	Rank	$R2LOG$	Rank	MAD_2	Rank	MAD_1	Rank	$HMSE$	Rank	SR	DA
GARCH-N	0.1167	6	1.0154	6	1.1552	7	0.3342	5	0.2697	5	0.6846	4	0.2335	9	0.79	11.8365**
GARCH-t	0.1177	7	1.0576	10	1.1532	5	0.3291	4	0.2679	4	0.6859	5	0.2311	8	0.80	12.2942**
GARCH-GED	0.1165	5	1.0384	9	1.1534	6	0.3281	3	0.2671	3	0.6816	3	0.2353	10	0.80	12.0885**
EGARCH-N	0.1288	8	0.8859	3	1.1772	8	0.4183	9	0.3106	9	0.7642	9	0.1946	5	0.81	13.4606**
EGARCH-t	0.1387	9	1.0288	8	1.2137	11	0.411	8	0.3003	8	0.7242	8	0.4977	13	0.67	5.0149**
EGARCH-GED	0.1608	11	1.249	11	1.1977	9	0.4769	10	0.3478	11	0.8903	11	0.2137	6	0.81	13.3399**
GJR-N	0.102	2	0.767	2	1.1442	2	0.3226	2	0.2616	2	0.6495	2	0.1667	2	0.83	13.8278**
GJR-t	0.1157	4	0.9453	5	1.1513	4	0.3439	7	0.2775	7	0.7071	7	0.1715	4	0.83	13.9998**
GJR-GED	0.111	3	0.8924	4	1.1482	3	0.335	6	0.2716	6	0.6879	6	0.169	3	0.83	13.8821**
MRS-GARCH-N	0.0686	1	0.4396	1	1.1192	1	0.2475	1	0.2111	1	0.4923	1	0.1544	1	0.79	12.4806**
MRS-GARCH-t2	0.1499	10	1.0193	7	1.2082	10	0.5074	11	0.3384	10	0.8161	10	0.2294	7	0.80	12.2593**
MRS-GARCH-t	0.226	13	1.5097	13	1.2749	13	0.7129	13	0.4266	13	1.0627	13	0.2832	12	0.81	12.6904**
MRS-GARCH-GED	0.1857	12	1.2767	12	1.2385	12	0.599	12	0.3832	12	0.9445	12	0.2537	11	0.81	12.6904**

5-step-ahead volatility forecasts

Model	MSE_1	Rank	MSE_2	Rank	$QLIKE$	Rank	$R2LOG$	Rank	MAD_2	Rank	MAD_1	Rank	$HMSE$	Rank	SR	DA
GARCH-N	0.476	7	21.4731	9	2.764	9	0.2579	7	0.536	7	3.0732	7	0.1619	9	0.78	12.0842**
GARCH-t	0.4813	9	22.6374	11	2.7619	6	0.2527	6	0.5314	6	3.0803	8	0.1587	7	0.79	12.4675**
GARCH-GED	0.4749	6	22.0788	10	2.762	7	0.2518	5	0.5301	5	3.0605	5	0.1609	8	0.79	12.2714**
EGARCH-N	0.3072	2	10.0325	2	2.7427	2	0.2115	2	0.4546	2	2.4296	2	0.1149	1	0.83	14.7704**
EGARCH-t	0.581	10	20.898	6	2.847	11	0.3235	10	0.5775	10	3.071	6	0.677	13	0.62	3.2527**
EGARCH-GED	0.4007	3	14.8289	3	2.7568	4	0.2489	4	0.5257	4	2.9322	4	0.1305	4	0.83	14.6525**
GJR-N	0.4096	4	16.3435	4	2.7558	3	0.2467	3	0.5127	3	2.8795	3	0.1298	3	0.82	13.8586**
GJR-t	0.4778	8	20.9307	7	2.7628	8	0.2665	9	0.5475	9	3.1745	10	0.1359	6	0.83	14.3333**
GJR-GED	0.4573	5	19.6661	5	2.7603	5	0.2596	8	0.5366	8	3.0878	9	0.1336	5	0.82	14.1382**
MRS-GARCH-N	0.2343	1	8.068	1	2.7274	1	0.1564	1	0.3769	1	1.9757	1	0.1277	2	0.80	12.8808**
MRS-GARCH-t2	0.6522	11	21.4315	8	2.8255	10	0.4372	11	0.7051	11	3.7787	11	0.2102	10	0.78	11.9647**
MRS-GARCH-t	0.9644	13	31.3951	13	2.8803	13	0.601	13	0.8826	13	4.8889	13	0.257	12	0.79	12.2077**
MRS-GARCH-GED	0.8133	12	28.8327	12	2.8484	12	0.504	12	0.7989	12	4.4342	12	0.2289	11	0.79	12.5770**

Note: The volatility proxy is given by the realized volatility calculated with one-minute returns and aggregated.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 6: Out-of-sample evaluation of the ten- and twenty-two-step-ahead volatility forecasts.

10-step-ahead volatility forecasts

Model	MSE_1	Rank	MSE_2	Rank	$QLIKE$	Rank	$R2LOG$	Rank	MAD_2	Rank	MAD_1	Rank	$HMSE$	Rank	SR	DA
GARCH-N	0.8886	8	77.5016	8	3.4565	9	0.2375	9	0.7541	10	6.1026	9	0.1551	8	0.78	11.7964**
GARCH-t	0.9014	9	82.4203	11	3.4543	7	0.2324	7	0.7496	9	6.1428	10	0.1516	6	0.79	12.2714**
GARCH-GED	0.8878	7	80.0378	10	3.4544	8	0.2315	6	0.7468	8	6.0888	8	0.1538	7	0.78	11.9817**
EGARCH-N	0.3281	1	22.3524	1	3.4052	1	0.1107	1	0.4416	1	3.2895	1	0.0815	1	0.83	14.4145**
EGARCH-t	1.061	10	75.3449	7	3.544	12	0.2844	10	0.7139	5	5.3417	4	0.7788	13	0.65	4.5233**
EGARCH-GED	0.4004	2	28.8078	2	3.4112	2	0.1278	2	0.4968	3	3.8117	3	0.085	2	0.83	14.5667**
GJR-N	0.725	4	56.5389	4	3.4446	4	0.2172	4	0.6875	4	5.4102	5	0.1204	3	0.80	13.1769**
GJR-t	0.8619	6	74.6869	6	3.4515	6	0.2369	8	0.7383	7	6.0228	7	0.126	5	0.81	13.4579**
GJR-GED	0.8249	5	69.9049	5	3.4496	5	0.2314	5	0.7244	6	5.8569	6	0.1243	4	0.81	13.3637**
MRS-GARCH-N	0.4448	3	32.4315	3	3.4233	3	0.1307	3	0.4938	2	3.6847	2	0.1594	9	0.78	12.1315**
MRS-GARCH-t2	1.2661	11	77.5618	9	3.5231	10	0.4286	11	1.0062	11	7.5573	11	0.2115	10	0.77	11.4764**
MRS-GARCH-t	1.7847	13	110.6103	12	3.5672	13	0.5611	13	1.2096	13	9.3906	13	0.2476	12	0.77	11.3732**
MRS-GARCH-GED	1.5652	12	111.2944	13	3.5396	11	0.4765	12	1.1242	12	8.8471	12	0.2241	11	0.79	12.1775**

22-step-ahead volatility forecasts

Model	MSE_1	Rank	MSE_2	Rank	$QLIKE$	Rank	$R2LOG$	Rank	MAD_2	Rank	MAD_1	Rank	$HMSE$	Rank	SR	DA
GARCH-N	1.8496	7	331.8337	7	4.2502	8	0.2219	9	1.1106	10	13.1113	9	0.1754	6	0.75	10.4164**
GARCH-t	1.8827	9	355.8608	11	4.2479	6	0.2168	8	1.1041	9	13.2158	10	0.1707	4	0.76	10.7708**
GARCH-GED	1.8536	8	344.544	8	4.2482	7	0.216	7	1.0996	8	13.0925	8	0.1748	5	0.76	10.7708**
EGARCH-N	1.0426	2	168.3615	2	4.2319	3	0.1302	2	0.7016	2	7.6866	2	0.2568	11	0.83	14.8844**
EGARCH-t	2.3655	10	345.6922	9	4.3616	13	0.3002	10	1.0587	7	11.4232	5	0.8939	13	0.63	6.6727**
EGARCH-GED	0.903	1	146.9349	1	4.2214	1	0.1156	1	0.652	1	7.1307	1	0.2148	7	0.83	14.8618**
GJR-N	1.3614	3	216.5264	3	4.2316	2	0.1876	3	0.9601	4	10.9613	4	0.1246	1	0.80	13.0751**
GJR-t	1.6322	6	292.4681	6	4.2373	5	0.2054	6	1.0292	6	12.1805	7	0.126	2	0.81	13.5560**
GJR-GED	1.5729	5	273.863	5	4.2365	4	0.2026	5	1.0176	5	11.9323	6	0.1261	3	0.81	13.4643**
MRS-GARCH-N	1.5118	4	232.1209	4	4.273	9	0.1884	4	0.8551	3	9.3355	3	0.433	12	0.77	11.7853**
MRS-GARCH-t2	2.7924	11	348.3369	10	4.3218	10	0.4306	11	1.5221	11	16.7397	11	0.226	9	0.77	10.9851**
MRS-GARCH-t	3.5822	13	453.4756	12	4.3526	12	0.5254	13	1.7272	13	19.4391	13	0.2484	10	0.74	9.9734**
MRS-GARCH-GED	3.2914	12	518.3628	13	4.3254	11	0.4433	12	1.6263	12	18.937	12	0.2235	8	0.77	11.1647**

Note: The volatility proxy is given by the realized volatility calculated with one-minute returns and aggregated.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 7 reports the DM test when the benchmark is the best model for the one-day horizon (MRS-GARCH-N), compared to each one of the other models. The comparison is carried out by taking into account all the statistical loss functions introduced in section 5.

It is evident that the MRS-GARCH-N (the benchmark) significantly outperforms every standard GARCH model at any usual confidence level. Remarkably, the sign of the DM statistic, when the benchmark is compared to standard GARCH models, is always negative, implying that the benchmark's loss is lower than the one implied by these models. When we consider the pairwise comparisons with the other MRS-GARCH models, we always reject the null of equal forecast accuracy. Only with the HMSE loss function we have some models for which we cannot reject the null hypothesis.[19]

Table 8, instead, presents the DM test when the benchmark model is the best at the two-week horizon (EGARCH-N). Here for all statistical loss functions and for all models but the MRS-GARCH-N we reject the null of EPA, suggesting that the benchmark fares the best. When the benchmark is compared to the second best (MRS-GARCH-N) we fail to reject the null of equal forecast accuracy for all loss functions but HMSE.

The results for all the other models and forecast horizons[20] show that when the benchmark is a GARCH model, tests of EPA are rejected for all MRS-GARCH but that one with normal innovations. In other words, the benchmark outperforms all MRS-GARCH but the MRS-GARCH-N which, in particular at shorter horizons, always implies a lower loss than the benchmark. In addition, EGARCH-N and EGARCH-GED also fare better than the benchmark at horizons longer than one day. When the benchmark is the EGARCH model, it beats almost all MRS-GARCH and some other standard GARCH. In particular, at shorter horizons (until one week) the MRS-GARCH-N always fares better than the bench-

[19]However, as shown in Patton (2005), the $HMSE$ loss does not imply that the conditional variance is the optimal forecast. Actually, the optimal forecast that minimizes this loss can be obtained by setting to zero the first derivative of the expected value of (5.7) w.r.t. $\widehat{h}_{t,t+m}$, which yields $\widehat{h}^{*}_{t,t+m} = E_{t-1}\left(\widehat{\sigma}^{4}_{t+m}\right)/E_{t-1}\left(\widehat{\sigma}^{2}_{t+m}\right)$ which is clearly different from the conditional variance $\widehat{\sigma}^{2}_{t+m}$, independently of the proxy used for the latent volatility. Therefore, the $HMSE$ loss is not particularly suitable for evaluating different volatility forecasts and it should be expected to give weird results.

[20]For all forecast horizons we have also computed the MDM statistics of Harvery, Leybourne and Newbold (1997). The overall results are only slightly different from the DM test and lead to exactly the same conclusions. This is due to the fact that the multiplicative factor $\sqrt{n^{-1}\left[n+1-2m+n^{-1}m(m-1)\right]}$ is .95, .98, .99, and .99 for the one-, five-, ten- and twenty-two-step-ahead horizon respectively. These results are also available upon request.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 7: Diebold-Mariano Test. (Benchmark: MRS-GARCH-N, Horizon: One day)

Model	MSE_1	MSE_2	$QLIKE$	$R2LOG$	MAD_2	MAD_1	$HMSE$
GARCH-N	-3.12**	-2.48*	-3.81**	-3.49**	-3.18**	-3.61**	-1.94
p-values	0.00	0.01	0.00	0.00	0.00	0.00	0.05
GARCH-t	-2.95**	-2.41*	-3.59**	-3.23**	-3.01**	-3.39**	-1.88
p-values	0.00	0.02	0.00	0.00	0.00	0.00	0.06
GARCH-GED	-2.96**	-2.42*	-3.60**	-3.23**	-3.02**	-3.40**	-1.87
p-values	0.00	0.02	0.00	0.00	0.00	0.00	0.06
EGARCH-N	-5.80**	-3.78**	-6.54**	-6.83**	-5.78**	-7.08**	-3.10**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EGARCH-t	-3.76**	-3.14**	-4.00**	-4.24**	-4.19**	-4.45**	-2.90**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EGARCH-GED	-6.16**	-3.99**	-7.71**	-7.92**	-6.18**	-7.83**	-4.23**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GJR-N	-3.79**	-2.63**	-4.30**	-4.90**	-3.87**	-4.53**	-1.18
p-values	0.00	0.01	0.00	0.00	0.00	0.00	0.24
GJR-t	-4.04**	-2.80**	-4.83**	-5.37**	-4.13**	-4.92**	-1.56
p-values	0.00	0.01	0.00	0.00	0.00	0.00	0.12
GJR-GED	-3.92**	-2.74**	-4.57**	-5.12**	-4.01**	-4.74**	-1.36
p-values	0.00	0.01	0.00	0.00	0.00	0.00	0.17
MRS-GARCH-t2	-7.44**	-3.92**	-8.23**	-8.18**	-8.10**	-9.81**	-5.76**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MRS-GARCH-t	-10.16**	-6.20**	-9.38**	-9.20**	-10.47**	-11.29**	-7.10**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MRS-GARCH-GED	-8.64**	-4.68**	-9.22**	-9.09**	-8.85**	-10.77**	-6.53**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: * and ** represent the DM statistics for which one can reject the null hypothesis of equal predictive accuracy at 5% and 1%, respectively.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 8: Diebold-Mariano Test. (Benchmark: EGARCH-N, Horizon: Two weeks)

Model	MSE_1	MSE_2	$QLIKE$	$R2LOG$	MAD_2	MAD_1	$HMSE$
GARCH-N	-4.47**	-3.00**	-7.35**	-6.69**	-4.96**	-6.51**	-8.05**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GARCH-t	-4.08**	-2.83**	-6.95**	-6.30**	-4.61**	-6.07**	-7.64**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GARCH-GED	-4.18**	-2.87**	-7.17**	-6.45**	-4.69**	-6.20**	-7.82**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EGARCH-t	-2.83**	-2.56*	-2.95**	-3.05**	-3.27**	-3.47**	-2.59**
p-values	0.00	0.01	0.00	0.00	0.00	0.00	0.01
EGARCH-GED	-2.74**	-2.09*	-2.98**	-3.83**	-2.99**	-3.32**	-0.82
p-values	0.01	0.04	0.00	0.00	0.00	0.00	0.41
GJR-N	-4.78**	-2.99**	-5.97**	-6.35**	-5.03**	-6.46**	-3.50**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GJR-t	-4.55**	-2.91**	-6.03**	-6.49**	-4.80**	-6.16**	-3.45**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GJR-GED	-4.58**	-2.91**	-6.01**	-6.46**	-4.82**	-6.22**	-3.44**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MRS-GARCH-N	-1.63	-1.59	-1.95	-1.20	-1.17	-1.28	-2.61**
p-values	0.10	0.11	0.05	0.23	0.24	0.20	0.01
MRS-GARCH-t2	-9.78**	-6.68**	-8.74**	-8.29**	-11.32**	-11.29**	-8.83**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MRS-GARCH-t	-9.95**	-8.38**	-8.91**	-8.42**	-11.87**	-11.58**	-8.88**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MRS-GARCH-GED	-8.60**	-4.72**	-9.69**	-9.27**	-9.55**	-11.83**	-8.88**
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: * and ** represent the DM statistics for which one can reject the null hypothesis of equal predictive accuracy at 5% and 1%, respectively.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

mark, whereas the other EGARCH models seem to outperform at longer horizons. Only EGARCH-t fails to reject the null of EPA with almost all other competing models. If GJR is the benchmark, it outperforms all MRS-GARCH but the MRS-GARCH-N and fares better than many other standard GARCH. The MRS-GARCH-N model outperforms at shorter horizons, while the EGARCH-N and EGARCH-GED have good performances at longer ones (more than one day). Furthermore, at all horizons, the GJR-N fares better than the same models with fat-tailed distributions. When the benchmark is the MRS-GARCH-N model, it outperforms all the other models until the one-week horizon, whereas it is beaten by the EGARCH-N and EGARCH-GED at longer horizons. If the benchmark is the MRS-GARCH-t2, it only outperforms the MRS-GARCH-t and MRS-GARCH-GED, but it is beaten by almost all the other models which imply a smaller loss. When the benchmark is the MRS-GARCH-t, all the other models outperform at all horizons since they significantly display a smaller loss than the benchmark. The same is true for the MRS-GARCH-GED which only beats the MRS-GARCH-t.

Therefore, we have seen that at shorter horizons MRS-GARCH-N fares the best, but at longer ones standard asymmetric GARCH models, such as the EGARCH-N, EGARCH-GED or GJR-N, tend to be superior. Another striking feature of all these pairwise analyses is that the other MRS-GARCH models with fat-tailed distributions are outperformed by almost all standard GARCH models. These results do not hold when the more general forecast evaluation for SPA is undertaken.

Table 9 reports the RC test for the SPA of each model against all the others at the one-day forecast horizon. The table presents for each benchmark model in the row and each loss function three p-values: the RC is the Reality Check p-value, while SPA_c^0 and SPA_l^0 are the Hansen's (2005) consistent and lower p-values, respectively.[21]

The p-values reported in Table 9 for the RC and SPA tests distinctly show how all these tests reject the null hypothesis of SPA when the benchmark is one of the standard GARCH models. This means that there is a competing model which is significantly better than the benchmark. This happens for all the single-regime GARCH models and for every loss function except for HMSE, for which

[21] Such p-values are calculated adopting the stationary bootstrap by Politis and Romano (1994) as in White (2000) and in Hansen and Lunde (2005). The number of bootstrap re-samples B is 3000 and the block length q is 0.33. However we have done the same calculations with $B = (1000, 2000, 3000)$ and a different set of values for the block lengths q (0.10, 0.20 and 0.33). The results do not change considerably. Therefore, we have chosen to report the table at the one-day horizon with $B = 3000$ and $q = 0.33$. The other tables are available upon request.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 9: Reality Check and SPA Tests (All models, Horizon: One day).

		Loss Functions						
Benchmark		MSE_1	MSE_2	$QLIKE$	$R2LOG$	MAD_1	MAD_2	$HMSE$
GARCH-N	SPA_l^0	0	0.003	0	0	0	0	0
	SPA_c^0	0	0.003	0	0	0	0.001	0
	RC	0.001	0.003	0.021	0.004	0	0.001	0.008
GARCH-t	SPA_l^0	0	0.002	0.001	0	0	0	0
	SPA_c^0	0	0.002	0.001	0	0	0	0
	RC	0.001	0.002	0.024	0.01	0	0	0.006
GARCH-GED	SPA_l^0	0	0.004	0	0	0	0	0
	SPA_c^0	0	0.004	0.001	0	0	0	0
	RC	0.001	0.004	0.022	0.013	0	0	0.007
EGARCH-N	SPA_l^0	0	0	0	0	0	0	0.036
	SPA_c^0	0	0	0	0	0	0	0.051
	RC	0	0	0	0	0	0	0.572
EGARCH-t	SPA_l^0	0	0.003	0	0	0	0	0
	SPA_c^0	0	0.004	0	0	0	0	0
	RC	0	0.004	0	0	0	0	0
EGARCH-GED	SPA_l^0	0	0	0	0	0	0	0.502
	SPA_c^0	0	0	0	0	0	0	0.725
	RC	0	0	0	0	0	0	0.939
GJR-N	SPA_l^0	0	0	0	0	0	0	0
	SPA_c^0	0	0.004	0.001	0	0	0	0
	RC	0.006	0.018	0.071	0.016	0	0.001	0.131
GJR-t	SPA_l^0	0	0.001	0	0	0	0	0
	SPA_c^0	0	0.001	0	0	0	0	0
	RC	0.002	0.002	0.03	0.002	0	0	0.174
GJR-GED	SPA_l^0	0	0	0	0	0	0	0
	SPA_c^0	0	0	0	0	0	0	0
	RC	0.001	0.003	0.044	0.003	0	0	0.148
MRS-GARCH-N	SPA_l^0	0.514	0.524	0.531	0.627	0.578	0.562	0
	SPA_c^0	1	1	1	1	1	1	0
	RC	1	1	1	1	1	1	0
MRS-GARCH-t2	SPA_l^0	0	0.001	0	0	0	0	0
	SPA_c^0	0	0.001	0	0	0	0	0
	RC	0	0.001	0	0	0	0	0.309
MRS-GARCH-t	SPA_l^0	0	0	0	0	0	0	0.492
	SPA_c^0	0	0	0	0	0	0	0.781
	RC	0	0	0	0	0	0	0.953
MRS-GARCH-GED	SPA_l^0	0	0	0	0	0	0	0.165
	SPA_c^0	0	0	0	0	0	0	0.302
	RC	0	0	0	0	0	0	0.837

Note: This table presents the p-values of White's (2000) Reality Check test (RC), and the p-values of the consistent (SPA_c^0) and lower bound (SPA_l^0) Hansen's (2005) Superior Predictive Ability (SPA) test for the one-step-ahead forecasts. Each model in the row is the benchmark versus all the other competitors. The null hypothesis is that none of the models is better than the benchmark. The number of bootstrap replications to calculate the p-values is 3000 and the block length is 0.33.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

EGARCH-N and EGARCH-GED are not beaten by another competing model. These apparently striking results are not new in the literature. Hansen and Lunde (2005) obtain similar results with stock market data, finding that the GARCH(1,1) specification is not the best model (in terms of SPA) when compared to other single-regime specifications. Table 9 also presents the RC and SPA test p-values when the benchmark is one of the MRS-GARCH models. It is evident that the MRS-GARCH-N model significantly outperforms all the other models at the usual significance level of 5%. As a matter of fact, for all the loss functions but $HMSE$ we fail to reject the null of no availability of a superior model. According to this loss function, EGARCH-N, EGARCH-GED, MRS-GARCH-t and MRS-GARCH-GED are the only models for which we cannot reject the null of SPA. Similar results are obtained for all forecast horizons and different block lengths. In general, MRS-GARCH-N is always the best model according to all the loss functions but HMSE, for which many other models fare the best. For some of the other loss functions and with shorter block lengths, we also find a few standard GARCH models that outperform all the competing models in terms of SPA.

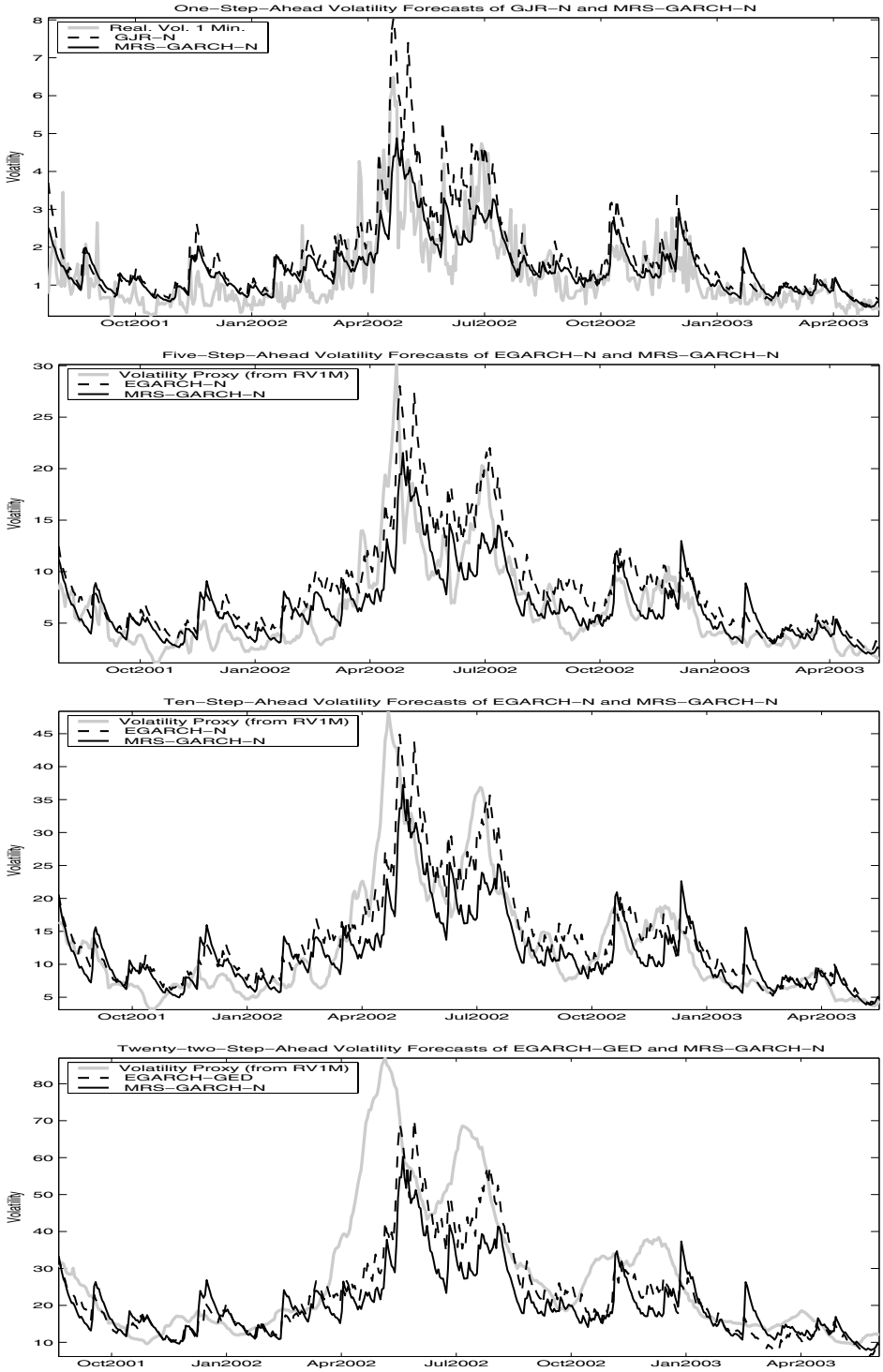
Table 10 reports the same RC and SPA test p-values when the comparison is made only among MRS-GARCH models at the two-week horizon. This table can therefore help us understand the possible implications of including poor models for these tests. The results are slightly different to the previous ones. Now, the MRS-GARCH-t significantly outperforms all the other MRS-GARCH, while the MRS-GARCH-GED fares the best according to MSE_2 and $QLIKE$ loss functions. We obtain quite similar results for different forecast horizons and shorter block lengths. The MRS-GARCH-t still outperforms all the other MRS-GARCH for every loss but HMSE and the MRS-GARCH-GED also fares the best according to some loss functions.

Figure 1 illustrates the volatility forecasts at the one-day, one-week, two-week and one-month horizons from the best models according to the out-of-sample evaluation based on statistical losses when the proxy for volatility is the realized volatility. Every sub-figure depicts the comparison between the forecasts of the best standard GARCH model and the best MRS-GARCH. From the plots it is evident that at shorter horizons standard GARCH models' volatility forecasts tend to have higher spikes than those of the MRS-GARCH, while the reverse is true at longer horizons. Thus, the model which fares the best gives much smoother forecasts and less spikes than the competitor.

These and the previous results must also be compared by means of a VaR-based evaluation criterion. Since one of the main purposes of volatility forecasting is to have an input for successive VaR estimation, it is necessary to see how

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Figure 1: Comparison of the one-, five-, ten- and twenty-two-step-ahead volatility forecasts from the best MRS-GARCH Model and the best GARCH.



Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 10: RC and SPA Tests (MRS-GARCH models, Horizon: Two weeks).

		Loss Functions						
Benchmark		MSE_1	MSE_2	$QLIKE$	$R2LOG$	MAD_1	MAD_2	$HMSE$
MRS-GARCH-N	SPA_l^0	0	0	0	0	0	0	0
	SPA_c^0	0	0	0	0	0	0	0
	RC	0	0	0	0	0	0	0
MRS-GARCH-t2	SPA_l^0	0	0.007	0.007	0	0	0	0
	SPA_c^0	0	0.007	0.007	0	0	0	0
	RC	0.001	0.007	0.007	0	0	0	0
MRS-GARCH-t	SPA_l^0	0.517	0.512	0.604	0.502	0.507	0.487	0
	SPA_c^0	0.517	1	0.604	0.502	1	1	0
	RC	1	1	1	1	1	1	0
MRS-GARCH-GED	SPA_l^0	0.001	0.014	0.06	0	0	0	0
	SPA_c^0	0.001	0.014	0.091	0	0	0	0
	RC	0.024	0.277	0.091	0.006	0.015	0.037	0

Note: This table presents the p-values of White's (2000) Reality Check test (RC), and the p-values of the consistent (SPA_c^0) and lower bound (SPA_l^0) Hansen's (2005) Superior Predictive Ability (SPA) test for the forecasts at the two-week horizon. Each model in the row is the benchmark versus all other MRS-GARCH models. The null hypothesis is that none of the models is better than the benchmark. The number of bootstrap replications to calculate the p-values is 3000 and the block length is 0.3.

competing models do fare in terms of a risk-management loss function.[22] This is closely related to the results of Dacco and Satchell (1999) who demonstrate that the evaluation of forecasts from non-linear models, such as regime-switching models, using statistical measures might be quite misleading. The authors propose to adopt alternative economic loss functions and their approach is followed by Brooks and Persand (2003) who use both statistical and risk-management loss functions to evaluate a set of models in terms of their ability to predict volatility. As already discussed in section 5, we go a little bit further with respect to Brooks and Persand (2003) by comparing the models in terms of both unconditional and conditional coverage of the corresponding VaR estimates. We also adopt Sarma *et al.*'s (2003) two-step evaluation procedure applying a second stage selection cri-

[22]Since all the statistical loss functions are of the MSE-type, using alternative approaches for forecast evaluation that convert the forecasts of (unobservable) volatility into forecasts of VaR might be useful. Actually, in this way we can have an 'indirect' evaluation of the underlying volatility forecasts in a context that is financially relevant. We would like to thank an anonymous referee for suggesting us this line of argument.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

terion of our VaR using subjective loss functions that should incorporate the risk manager's preferences. These loss functions (RLF and FLF) take into account the magnitude of the failures in the VaR forecast, penalizing more the bigger ones. In addition, FLF penalizes those models that require a too high amount of capital by taking into account the opportunity cost of a firm to have too much capital tied up.

Tables 11 and 12 report the risk-management out-of-sample first-step evaluation of our competing GARCH models in terms of the 95% and 99% VaR, respectively, for the one-day, one-week, two-week and one-month horizons.

Seven statistics are presented for each model and each forecast horizon: the $TUFF$, the PF, the test of correct unconditional coverage (LR_{PF}) to check if PF is significantly higher than the nominal rate, the LR_{IND} for independence, the test of correct conditional coverage LR_{CC}, and the average RLF and FLF.

The theoretical $TUFF$ at 5 and 1% should be 20 and 100, respectively. We can thus see that only for the 99% VaR at the one-day horizon some models fail before the theoretical TUFF in addition to the EGARCH-N and the MRS-GARCH-N at twenty-two days. In all the other cases, failures happen after the theoretical level. If we instead look at the 95% confidence interval for the $TUFF$ (or equivalently to the significant LR_{TUFF} not reported in the table since they give the same result), we notice that we have some rejections only for the 95% VaR. In particular, we reject the null for both MRS-GARCH-t2 and MRS-GARCH-t at the five-, ten-, and twenty-two-day horizons, as well as for all the standard GARCH models with Student's t innovations at the two-week horizon, and for the EGARCH-t at the one-month horizon. However, this statistic is known to have low power and can be highly sensitive to the degree of turbulence at the beginning of the hold-out sample.

If the objective is to cover either the 99% or 95% of future losses, then many models seem inadequate, especially at the longest forecast horizons. In fact, most models display a PF below the nominal one but only at the shortest horizons. It is noticeable that at all horizons and for both coverage probabilities the best model according to the statistical forecast evaluation criteria - i.e. the MRS-GARCH-N - is always rejected for a too high PF.

The three tests of correct unconditional and conditional coverage show that we have a clear-cut answer only at the one-day forecast horizon, where all MRS-GARCH models but the MRS-GARCH-GED lack some of these properties. At all the other horizons, every model is rejected for not having the correct coverage either conditionally or unconditionally. The numbers in boldface represent the minima within each group of models (standard GARCH and MRS-GARCH) and help us to select the best models in terms of unconditional and conditional

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 11: Risk-management Out-of-sample Evaluation: 95% VaR

95% VaR														
Steps	**1**							**5**						
Model	TUFF	PF(%)	**LRPF**	**LRind**	**LRcc**	**RLF**	**FLF**	TUFF	PF(%)	**LRPF**	**LRind**	**LRcc**	**RLF**	**FLF**
GARCH-N	20	6.458	2.102	0.010	2.112	0.054	0.082	49	5.871	0.775	34.293*	35.067*	0.209	0.273
GARCH-t	20	3.718	1.932	0.116	**2.049**	0.027	0.062	86	2.544	7.854*	21.812*	29.665*	0.107	0.186
GARCH-GED	20	6.458	2.102	0.010	2.112	0.053	0.081	49	5.871	0.775	34.293*	35.067*	0.203	0.267
EGARCH-N	20	4.892	**0.013**	2.579	2.591	0.032	0.063	49	4.892	**0.013**	26.224*	26.236*	0.173	0.238
EGARCH-t	20	4.697	0.101	2.443	2.544	0.038	0.070	49	4.892	**0.013**	32.066*	32.079*	0.354	0.420
EGARCH-GED	20	3.718	1.932	1.471	3.403	0.028	0.062	49	4.305	0.544	25.474*	26.018*	0.155	0.222
GJR-N	20	5.675	0.471	3.499	3.970	0.045	0.075	49	4.892	**0.013**	20.859*	20.871*	0.159	0.225
GJR-t	20	3.718	1.932	1.471	3.403	**0.023**	**0.060**	49	2.544	7.854*	4.319*	**12.173***	**0.069**	**0.151**
GJR-GED	20	5.479	0.240	3.255	3.495	0.043	0.074	49	4.501	0.277	18.460*	18.737*	0.148	0.216
MRS-GARCH-N	20	7.045	4.014*	0.092	4.106	0.061	0.088	41	7.241	4.773*	32.357*	**37.130***	0.288	0.345
MRS-GARCH-t2	86	1.761	14.876*	2.164	17.041*	**0.009**	0.054	200	0.978	25.646*	23.353*	48.999*	**0.084**	**0.185**
MRS-GARCH-t	20	2.544	7.854*	0.945	8.798*	0.012	**0.053**	187	1.37	19.674*	17.739*	37.413*	0.105	0.194
MRS-GARCH-GED	20	4.305	**0.544**	0.989	**1.533**	0.035	0.067	49	3.914	**1.367**	35.952*	37.319*	0.191	0.261

95% VaR														
Steps	**10**							**22**						
Model	TUFF	PF(%)	**LRPF**	**LRind**	**LRcc**	**RLF**	**FLF**	TUFF	PF(%)	**LRPF**	**LRind**	**LRcc**	**RLF**	**FLF**
GARCH-N	77	5.088	**0.008**	42.734*	42.743*	0.619	0.710	69	9.002	14.070*	135.694*	149.764*	1.201	1.328
GARCH-t	186	1.566	17.148*	24.637*	41.784*	0.394	0.506	70	4.501	**0.277**	76.168*	76.445*	0.636	0.795
GARCH-GED	77	5.284	0.085	47.098*	47.183*	0.610	0.700	69	9.002	14.070*	135.694*	149.764*	1.197	1.323
EGARCH-N	77	5.479	0.240	44.659*	44.899*	0.738	0.822	65	13.894	58.625*	156.722*	215.347*	2.157	2.252
EGARCH-t	169	4.892	0.013	60.188*	60.201*	0.832	0.921	129	8.806	12.832*	131.663*	144.496*	1.696	1.810
EGARCH-GED	77	4.892	0.013	38.380*	38.392*	0.673	0.760	65	12.916	47.839*	141.262*	189.101*	1.966	2.065
GJR-N	77	3.718	1.932	24.513*	**26.445***	0.623	0.717	69	8.806	12.832*	114.588*	127.420*	1.105	1.233
GJR-t	187	1.761	14.876*	22.123*	36.999*	**0.373**	**0.488**	70	4.11	0.906	48.444*	**49.350***	**0.486**	**0.648**
GJR-GED	77	3.718	1.932	24.513*	**26.445***	0.583	0.678	69	8.611	11.643*	118.947*	130.590*	1.006	1.137
MRS-GARCH-N	70	8.611	11.643*	67.578*	79.221*	0.997	1.070	65	17.417	103.924*	261.995*	365.918*	2.937	3.023
MRS-GARCH-t2	200	0.587	33.279*	18.523*	51.802*	**0.298**	**0.438**	175	1.761	14.876*	42.362*	**57.239***	**0.307**	**0.506**
MRS-GARCH-t	187	1.761	14.876*	14.026*	**28.903***	0.396	0.517	175	3.327	**3.397**	62.162*	65.559*	0.566	0.731
MRS-GARCH-GED	77	4.11	**0.906**	33.734*	34.640*	0.599	0.695	69	8.415	10.503*	114.918*	125.422*	1.178	1.308

Note: This table presents the time until first failure (TUFF), the percentage proportion of failures (PF(%)), the likelihood ratio (LR) test for unconditional coverage (**LRpf**), the LR test for independence (**LRind**), the LR test for conditional coverage (**LRcc**), and the average regulator (RLF) and firm loss function (FLF) for the 95% VaR failure processes at one-, five-, ten- and twenty-two-step-ahead. Numbers in boldface are the minima of each group (standard GARCH and MRS-GARCH). * indicates significance at 5%.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 12: Risk-management Out-of-sample Evaluation: 99% VaR

99% VaR														
Steps	**1**							**5**						
Model	TUFF	PF(%)	**LRPF**	**LRind**	**LRcc**	**RLF**	**FLF**	TUFF	PF(%)	**LRPF**	**LRind**	**LRcc**	**RLF**	**FLF**
GARCH-N	20	2.544	8.621*	0.945	9.566*	0.011	0.053	206	0.587	1.033	6.844*	7.878*	0.059	0.154
GARCH-t	86	0.391	2.487	0.016	2.503	0.001	0.060	207	0.196	4.991*	0.004	4.995	0.004	0.136
GARCH-GED	86	1.174	**0.148**	0.143	**0.291**	0.005	0.053	206	0.391	2.487	8.926*	11.412*	0.033	0.141
EGARCH-N	20	1.566	1.408	0.255	1.663	0.005	**0.051**	187	0.783	0.263	5.505*	5.768	0.035	0.131
EGARCH-t	86	0.783	0.263	0.063	0.326	0.002	0.056	200	1.174	0.148	20.197*	20.346*	0.097	0.209
EGARCH-GED	86	0.587	1.033	0.036	1.069	0.001	0.056	206	0.391	2.487	8.926*	11.412*	0.009	0.121
GJR-N	20	1.761	2.438	0.323	2.762	0.011	0.055	187	0.978	**0.002**	4.522*	**4.525**	0.027	0.125
GJR-t	86	0.587	1.033	0.036	1.069	**0.000**	0.061	511	0	10.271*	0.000	10.271*	**0.000**	0.135
GJR-GED	20	1.37	0.633	0.195	0.828	0.005	0.055	206	0.391	2.487	8.926*	11.412*	0.006	**0.118**
MRS-GARCH-N	20	2.348	6.803*	0.578	7.382*	0.015	**0.055**	187	1.566	**1.408**	15.727*	17.135*	0.078	0.163
MRS-GARCH-t2	511	0	10.271*	0.000	10.271*	**0.000**	0.074	207	0.196	4.991*	0.004	**4.995**	**0.002**	0.165
MRS-GARCH-t	511	0	10.271*	0.000	10.271*	**0.000**	0.069	207	0.196	4.991*	0.004	**4.995**	**0.002**	0.152
MRS-GARCH-GED	86	0.783	**0.263**	0.063	**0.326**	0.001	**0.055**	206	0.391	2.487	8.926*	11.412*	0.025	**0.144**

99% VaR														
Steps	**10**							**22**						
Model	TUFF	PF(%)	**LRPF**	**LRind**	**LRcc**	**RLF**	**FLF**	TUFF	PF(%)	**LRPF**	**LRind**	**LRcc**	**RLF**	**FLF**
GARCH-N	187	0.783	0.263	15.083*	15.346*	0.248	0.382	175	1.761	2.438	42.362*	44.801*	0.339	0.532
GARCH-t	200	0.587	1.033	18.523*	19.556*	0.033	0.217	188	0.587	1.033	18.523*	19.556*	0.035	0.304
GARCH-GED	200	0.587	1.033	18.523*	19.556*	0.152	0.303	187	0.978	**0.002**	23.353*	23.355*	0.185	0.405
EGARCH-N	187	1.566	1.408	24.637*	26.045*	0.296	0.420	70	5.871	57.665*	92.787*	150.453*	0.610	0.757
EGARCH-t	187	1.174	0.148	10.944*	**11.093***	0.236	0.385	187	1.761	2.438	42.362*	44.801*	0.222	0.421
EGARCH-GED	200	0.978	**0.002**	23.353*	23.355*	0.162	0.306	129	2.153	5.156*	56.608*	61.764*	0.312	0.488
GJR-N	187	1.37	0.633	17.739*	18.372*	0.226	0.362	175	1.761	2.438	42.362*	44.801*	0.218	0.413
GJR-t	200	0.587	1.033	18.523*	19.556*	**0.018**	**0.206**	189	0.196	4.991*	0.004	**4.995**	**0.001**	**0.272**
GJR-GED	200	0.978	**0.002**	23.353*	23.355*	0.101	0.257	187	1.174	0.148	31.493*	31.641*	0.067	0.292
MRS-GARCH-N	186	2.153	5.156*	26.028*	31.185*	0.458	0.570	69	7.828	97.298*	119.832*	217.130*	0.946	1.084
MRS-GARCH-t2	200	0.587	**1.033**	18.523*	**19.556***	**0.008**	0.235	511	0	10.271*	0.000	**10.271***	**0.000**	0.332
MRS-GARCH-t	200	0.587	**1.033**	18.523*	**19.556***	0.027	**0.232**	189	0.391	2.487	8.926*	11.412*	0.005	**0.296**
MRS-GARCH-GED	200	0.587	**1.033**	18.523*	**19.556***	0.136	0.301	187	0.783	**0.263**	27.804*	28.067*	0.138	0.371

Note: This table presents the time until first failure (TUFF), the percentage proportion of failures (PF(%)), the likelihood ratio (LR) test for unconditional coverage (**LRpf**), the LR test for independence (**LRind**), the LR test for conditional coverage (**LRcc**), and the average regulator (RLF) and firm loss function (FLF) for the 99% VaR failure processes at one-, five-, ten- and twenty-two-step-ahead. Numbers in boldface are the minima of each group (standard GARCH and MRS-GARCH). * indicates significance at 5%.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

coverage.[23] Tables 11 and 12 also display the average of RLF and FLF. GJR-t and MRS-GARCH-t2 are the models that more frequently give the lowest averages of both losses. These models are also chosen for the second-stage selection criterion and are compared with the best ones according to conditional or unconditional coverage.

Tables 13 and 14 report the sign tests for the pairs of best models according to the two procedures discussed above. The most striking result is that under the RLF at all horizons and for all coverage probabilities, none of the models is significantly better than the others. This happens when we compare models with the lowest averages and also when the comparison is among models with the best unconditional or conditional coverage. Under the FLF, the results are completely different, meaning that it is important to take into account the opportunity cost of capital. For the 95% VaR, among the models with the lowest average FLF, the GJR-t is always significantly better than its MRS-GARCH competitor. On the other hand, among the models selected in terms of correct conditional or unconditional coverage, the best models are the EGARCH-N at the one-day horizon, the EGARCH-N and GJR-N at the one-week horizon, the GARCH-N at the two-week horizon and the GARCH-t at the one-month horizon. For the 99% VaR the best models among those pairs with the lowest averages are the MRS-GARCH-N, the GJR-GED, the GJR-t and again the GJR-t, respectively at the forecast horizons of one day, one week, two weeks and one month. Conversely, the best models, among those selected through the correct coverage, are the GJR-t, the MRS-GARCH-N, the EGARCH-GED and the GARCH-GED, at the one-day, one-week, two-week and one-month forecast horizon, respectively.

From all these results, it is not clear which model is the uniformly most accurate. Overall, there is some evidence that the GJR-t outperforms the competitors more often.

Nevertheless, in two cases the MRS-GARCH-N fares the best and in particular at the shortest horizons (one day and one week), where it also beats all the competitors under most statistical loss functions of the MSE-type. At horizons longer than one week, other standard asymmetric GARCH models seems preferable in terms of risk-management loss functions. This result is however not new, since also Brooks and Persand (2003) find a no clear answer for most of the series

[23] We should thus consider the second stage of selection only for the one-day horizon. However, for completeness, we also apply such selection to all the other horizons, selecting the best models among those that exhibit the correct coverage at least in one case. For example, for the 99% VaR at the one-week horizon, among the MRS-GARCH we select the MRS-GARCH-N under LR_{PF} and MRS-GARCH-t and MRS-GARCH-t2 under LR_{CC}.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 13: Sign tests of RLF and FLF: 95% VaR

95% VaR												
Steps	**1**						**5**					
	RLF			**FLF**			**RLF**			**FLF**		
	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$
Lowest Average	(8,11)	-21.013*	-22.428*	(8,12)	-11.457*	11.457	(8,11)	-21.809*	-22.163*	(8,11)	-18.093*	18.093
Best UC and CC	(4,13)	-21.190*	-21.455*	(4,13)	-6.326*	6.326	(7,13)	-21.101*	-21.632*	(7,13)	-8.361*	8.361
	(2,13)	-21.986*	-21.190*	(2,13)	5.441	-5.441*	(4,13)	-21.101*	-21.721*	(4,13)	-10.484*	10.484
	(2,4)	-21.809*	-21.013*	(2,4)	9.511	-9.511*	(5,13)	-20.570*	-21.809*	(5,13)	-1.017	1.017
							(4,5)	-21.190*	-21.013*	(4,5)	-7.211*	7.211
							(4,7)	-21.278*	-21.544*	(4,7)	-1.46	1.46
							(5,7)	-21.013*	-21.190*	(5,7)	5.795	-5.795*
95% VaR												
Steps	**10**						**22**					
	RLF			**FLF**			**RLF**			**FLF**		
	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$
Lowest Average	(8,11)	-21.986*	-22.428*	(8,11)	-17.562*	17.562	(8,11)	-21.013*	-22.340*	(8,11)	-16.943*	16.943
Best UC and CC	(1,13)	-20.924*	-21.986*	(1,13)	-11.988*	11.988	(2,12)	-21.013*	-21.898*	(2,12)	-6.768*	6.768

Note: This table presents the sign tests on loss differentials between pairs of best models according to the first-step of the risk-management out-of-sample selection procedure. These best models are selected either from those with the lowest average of the RLF or FLF in each group or from those which pass the coverage tests in the first step. $\widehat{S}_{ij} = (S_{ij} - 0.5T)(0.25T)^{-.5}$, where $S_{ij} = \sum_{t=1}^{T} \mathcal{I}_{\{z_t>0\}}$, $\mathcal{I}_{\{\}}$ is the indicator function and $z_t = l_{it}^k - l_{jt}^k$. $\widehat{S}_{ij}$ is asymptotically distributed as a standard normal. * indicates significance at 5%. A rejections of $\widehat{S}_{ij}$ means that model i is significantly better than model j and viceversa for $\widehat{S}_{ji}$.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Table 14: Sign tests of RLF and FLF: 99% VaR

99% VaR												
Steps	**1**						**5**					
	RLF			**FLF**			**RLF**			**FLF**		
	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$
Lowest Average	(8,11)	-22.340*	-22.605*	(4,13)	-18.182*	18.182	(8,11)	-22.605*	-22.517*	(9,13)	-10.042*	10.042
	(8,12)	-22.340*	-22.605*	(4,10)	16.412	-16.412*	(8,12)	-22.605*	-22.517*			
	(11,12)	-22.605*	-22.605*	(10,13)	-21.721*	21.721	(11,12)	-22.605*	-22.517*			
Best UC and CC	(3,13)	-22.074*	-22.605*	(3,13)	-17.562*	17.562	(7,10)	-22.605*	-21.898*	(7,10)	18.447	-18.447*
							(7,11)	-22.163*	-22.605*	(7,11)	-22.251*	22.251
							(7,12)	-22.163*	-22.605*	(7,12)	-21.544*	21.544
							(10,11)	-21.898*	-22.605*	(10,11)	-22.163*	22.163
							(10,12)	-21.898*	-22.605*	(10,12)	-22.163*	22.163
							(11,12)	-22.605*	-22.517*	(11,12)	16.854	-16.854*

99% VaR												
Steps	**10**						**22**					
	RLF			**FLF**			**RLF**			**FLF**		
	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$	(i,j)	$\widehat{S}_{ij}$	$\widehat{S}_{ji}$
Lowest Average	(8,11)	-22.517*	-22.428*	(8,12)	-11.369*	11.369	(8,11)	-22.517*	-22.605*	(8,12)	-10.219*	10.219
Best UC and CC	(6,11)	-22.163*	-22.605*	(6,11)	-22.163*	22.163	(3,13)	-22.163*	-22.605*	(3,13)	-13.935*	13.935
	(6,12)	-22.163*	-22.605*	(6,12)	-21.367*	21.367	(8,13)	-22.605*	-22.251*	(8,13)	14.643	-14.643*
	(6,13)	-22.251*	-22.517*	(6,13)	-16.677*	16.677	(3,8)	-22.163*	-22.605*	(3,8)	-21.013*	21.013
	(9,11)	-22.163*	-22.605*	(9,11)	-21.544*	21.544						
	(9,12)	-22.163*	-22.605*	(9,12)	-19.597*	19.597						
	(9,13)	-22.340*	-22.428*	(9,13)	-8.892*	8.892						
	(6,9)	-22.163*	-22.605*	(6,9)	-16.501*	16.501						
	(11,12)	-22.605*	-22.340*	(11,12)	18.889	-18.889*						
	(11,13)	-22.605*	-22.340*	(11,13)	22.34	-22.340*						
	(12,13)	-22.605*	-22.340*	(12,13)	20.747	-20.747*						

Note: This table presents the sign tests on loss differentials between pairs of best models according to the first-step of the risk-management out-of-sample selection procedure. These best models are selected either from those with the lowest average of RLF or FLF in each group or from those which pass the coverage tests in the first step. $\widehat{S}_{ij} = (S_{ij} - 0.5T)(0.25T)^{-.5}$, where $S_{ij} = \sum_{t=1}^{T} \mathcal{I}_{\{z_t>0\}}$, $\mathcal{I}_{\{\}}$ is the indicator function and $z_t = l_{it}^k - l_{jt}^k$. $\widehat{S}_{ij}$ is asymptotically distributed as a standard normal. * indicates significance at 5%. A rejections of $\widehat{S}_{ij}$ means that model i is significantly better than model j and viceversa for $\widehat{S}_{ji}$.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

they examine.

Figures 2 and 3 depict the excessive losses of the 95% and 99% VaR from the GJR-t and MRS-GARCH-t2 models. There are some differences between the two models so that the MRS-GARCH-t2 seems slightly worse than the GJR-t to capture the quick changes in the volatility of the returns. In a certain sense, the VaR from MRS-GARCH-t2 looks somehow smoother than that of GJR-t and this could be the result of the averaging between the two regimes.

In sum, none of the models seems to outperform all the others in forecasting the US stock market volatility at all horizons according to the different out-of-sample evaluation criteria adopted. Accounting for regime shifts in all the parameters of the first two conditional moments with GARCH effects and fat-tailed-ness, we obtain a better in-sample fit and outstanding out-of-sample results in terms of both EPA and SPA under the usual statistical loss functions of the MSE-type. However, at shorter horizons of one day and one week the MRS-GARCH-N, that outperforms all the others under statistical losses, also fares the best in terms of FLF, a more realistic risk-management loss.

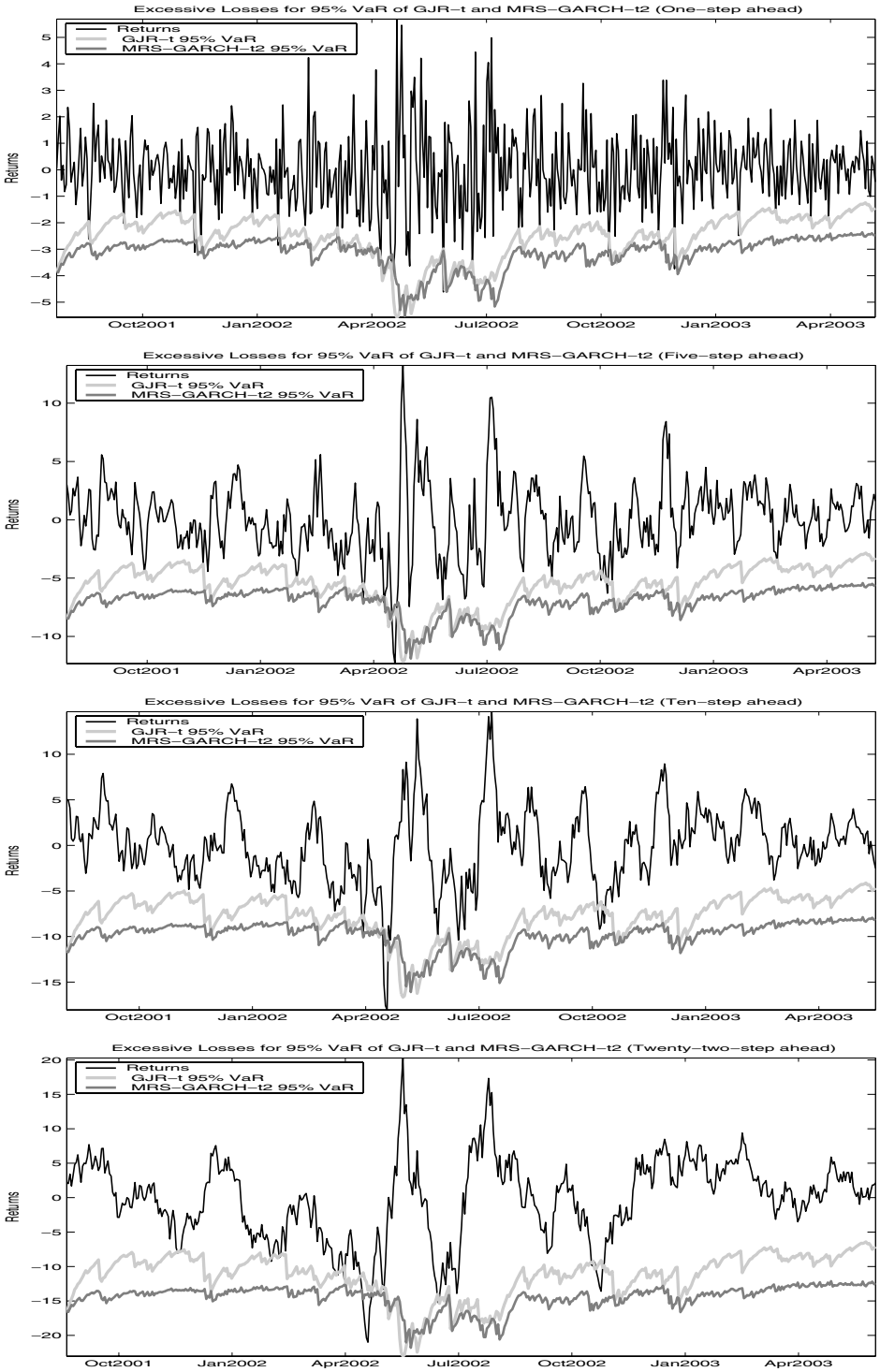
These findings partially agree with Dacco and Satchell's (1999) arguments that the choice of the correct loss function is fundamental for the accuracy of forecasts from non-linear models. However, our results do not confirm Dacco and Satchell's (1999) theoretical findings that although most non-linear techniques give a good in-sample fit, they are usually outperformed in out-of-sample forecasting by simpler models using economic loss functions. They argue that such a typical finding may be due to possible over-fitting and to the MSE metric that might be inappropriate for non-linear models.

Nevertheless, even though our results from the out-of-sample evaluation in terms of risk-management losses do not give a clear-cut answer, we have some evidence that MRS-GARCH models do fare better than any other simpler competitors, even when we use more realistic economic loss functions.

Therefore, three main results have emerged from the present work. Firstly, at short horizons (from one day to one week) regime-switching models with GARCH effects do really outperform standard GARCH in predicting volatility under both statistical and risk-management loss functions. Secondly, at horizons longer than one week, standard asymmetric GARCH models fare better than MRS-GARCH using both statistical and VaR-based losses. Finally, under risk-management loss functions, taking into account the opportunity cost of capital along with the magnitude of the VaR failures is crucial to select the best model.

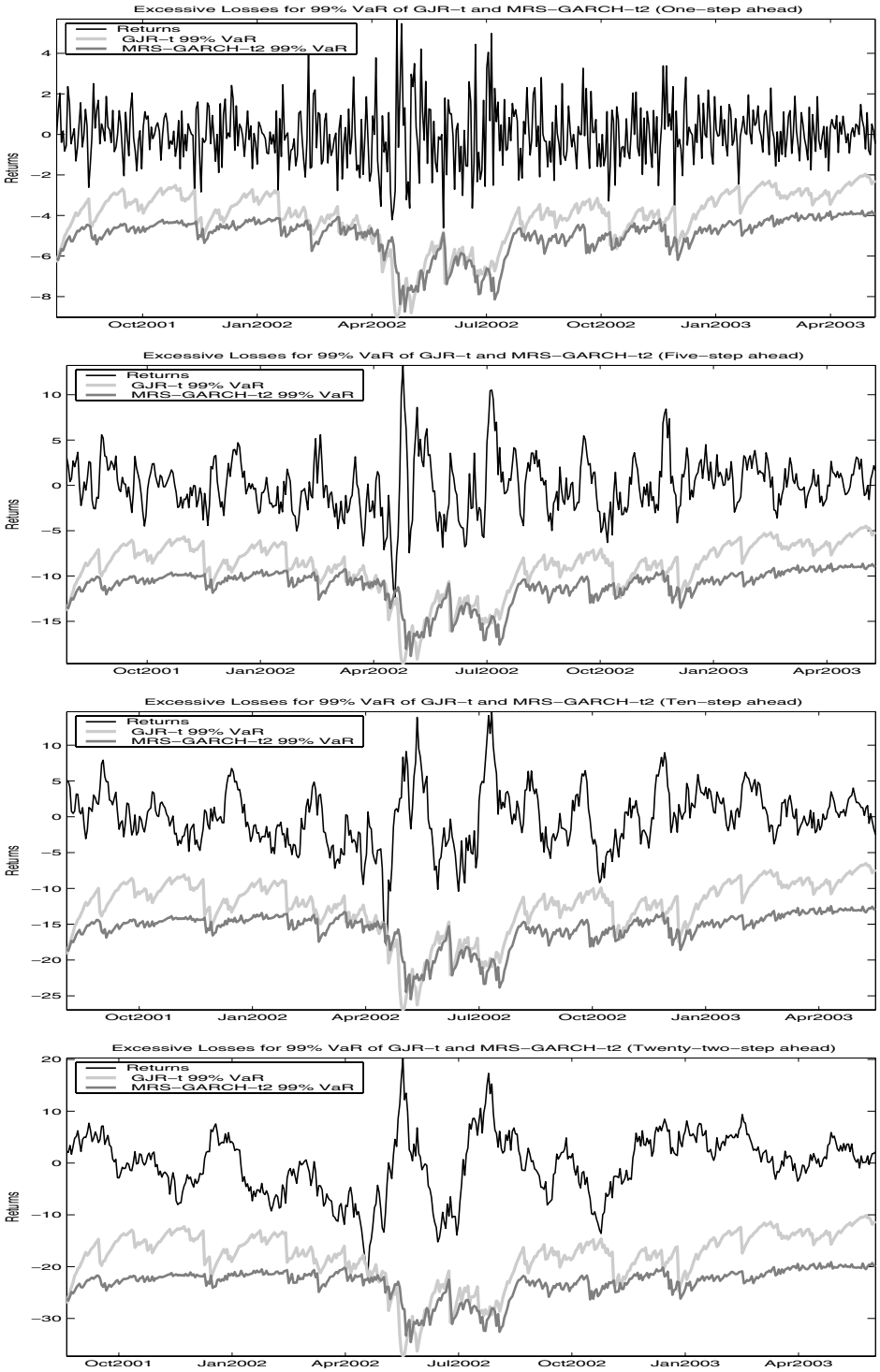
Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Figure 2: 95% VaR estimates from the best models.



Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Figure 3: 99% VaR estimates from the best models.



Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

7 Conclusions

In this paper we compare a set of standard GARCH models and Markov Regime-Switching GARCH (MRS-GARCH) in terms of their ability to forecast US stock market volatility. We employ standard GARCH models such as the GARCH(1,1), EGARCH(1,1) and GJR(1,1), in addition to some MRS-GARCH models, where each parameter of the first two conditional moments switches between two regimes with different volatilities. Moreover, all the models are estimated assuming both gaussian innovations and fat-tailed distributions, such as the Student's t and the GED. Further, to capture possible time-varying conditional kurtosis, the degrees-of-freedom parameter in the Student's t distribution can switch across the two regimes in a completely different setting than those considered by Hansen (1994) and Dueker (1997).

The main goal is to evaluate the performance of different GARCH models in terms of their ability to characterize and predict out-of-sample the volatility of the S&P100 index. The out-of-sample evaluation is carried out by comparing the volatility forecasts from all models over the one-day, one-week, two-week and one-month horizons.

As a proxy for volatility we use the realized volatility calculated as the sum of one-minute squared returns aggregated over the relevant forecast horizon. The forecasting performances of each model are evaluated using both statistical and Value-at-Risk-based risk-management loss functions. Under statistical losses, we compare standard GARCH and MRS-GARCH in terms of both equal predictive ability (EPA) with tests of the Diebold-Mariano-type and superior predictive ability (SPA) with the Reality Check of White (2000) and the SPA test of Hansen (2005). Since volatility forecasts are mainly used as an input for successive VaR estimation, we also evaluate the competing models out-of-sample by means of risk-management loss functions. Under these losses, we adopt a two-step selection procedure, where in the first stage we test the competing models in terms of correct conditional or unconditional coverage of their VaR's, identifying those models that pass such tests as the best. In the second stage, we compare pairs of best models under two VaR-based economic loss functions to find the model that significantly outperforms the others. These economic losses are the regulator loss function and the firm loss function (FLF). While both losses take into account the magnitude of the failures in the VaR, the FLF also considers the opportunity cost of capital.

Overall, the empirical results show that the MRS-GARCH model with normal innovations significantly outperforms standard GARCH models and the other

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

MRS-GARCH in forecasting volatility at horizons shorter than one week both under statistical and risk-management losses. Standard asymmetric GARCH fare better at longer horizons. However, our results show that in terms of risk-management losses we do not have a clear-cut answer. This result is not new, since also Brooks and Persand (2003) find a no clear answer for most of the series they examine. Nevertheless, for the 99% VaR at forecast horizons shorter than one week, the MRS-GARCH model with normal innovations fares the best also under these risk-management losses.

In sum, none of the models seems to be uniformly superior in forecasting the US stock market volatility. Accounting for regime shifts in all the parameters of the first two moments of the conditional distribution with GARCH effects and fat-tailed-ness, gives a better in-sample fit and outstanding out-of-sample results in terms of both equal and superior predictive ability under the usual statistical loss functions of the MSE-type. However, at shorter horizons of one day and one week the MRS-GARCH with normal innovations, that outperforms all the others under statistical losses, also fares the best in terms of more realistic risk-management losses. These results do not confirm Dacco and Satchell's (1999) theoretical findings that although most non-linear techniques give good in-sample fit, they are usually outperformed in out-of-sample forecasting by simpler models using economic loss functions. They argue that such a typical finding may be due to possible over-fitting and to the mean squared error metric that might be inappropriate for highly non-linear models. Nevertheless, even though our results from the out-of-sample evaluation in terms of risk-management losses do not give a clear-cut answer, we have some evidence that MRS-GARCH models do fare better than other non-linear competitors even when we use more realistic VaR-based economic loss functions.

Therefore, three main results have emerged from this paper. Firstly, at short horizons (from one day to one week) regime-switching models with GARCH effects do really outperform standard GARCH in predicting volatility under both statistical and risk-management loss functions. Secondly, at horizons longer than one week, standard asymmetric GARCH models fare better than MRS-GARCH using both statistical and VaR-based losses. Finally, under risk-management loss functions, taking into account the opportunity cost of capital along with the magnitude of the VaR failures is crucial to select the best model.

Further research is however needed to build more parsimonious volatility models and to construct *ad hoc* out-of-sample evaluation procedures to correctly identify the best models in terms of volatility forecasting.

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

References

Andersen, T. G., and T. Bollerslev (1998) 'Answering the Critics: Yes ARCH Models Do Provide Good Volatility Forecasts.' *International Economic Review* 39(4), 885–905

Billio, M., and L. Pellizzon (2000) 'Value-at-Risk: A Multivariate Switching Regime Approach.' *Journal of Empirical Finance* 7(5), 531–554

Bollerslev, T. (1986) 'Generalized Autoregressive Conditional Heteroskedasticity.' *Journal of Econometrics* 31(3), 307–327

Bollerslev, T., and E. Ghysels (1996) 'Periodic Autoregressive Conditional Heteroskedasticity.' *Journal of Business and Economic Statistics* 14(2), 139–151

Bollerslev, T., R. F. Engle, and D. Nelson (1994) 'ARCH Models.' In *Handbook of Econometrics Vol. IV,* ed. R. F. Engle and D. L. McFadden (Amsterdam: North-Holland) pp. 2959–3038

Brooks, C., and G. Persand (2003) 'Volatility Forecasting for Risk Management.' *Journal of Forecasting* 22(1), 1–22

Cai, J. (1994) 'A Markov Model of Unconditional Variance in ARCH.' *Journal of Business and Economic Statistics* 12(3), 309–316

Calvet, L. E. and A. J. Fisher (2004) 'How to Forecast Long-Run Volatility: Regime-Switching and the Estimation of Multifractal Processes.' *Journal of Financial Econometrics* 2(1), 49–83

Christoffersen, P. F. (1998) 'Evaluating Interval Forecasts.' *International Economic Review* 39(4), 841–862

Dacco, R., and S. Satchell (1999) 'Why Do Regime-Switching Models Forecast so Badly?' *Journal of Forecasting* 18(1), 1–16

Diebold, F. X., and R. S. Mariano (1995) 'Comparing Predictive Accuracy.' *Journal of Business and Economic Statistics* 13(3), 253–263

Dueker, M. J. (1997) 'Markov Switching in GARCH Processes and Mean-Reverting Stock Market Volatility.' *Journal of Business and Economic Statistics* 15(1), 26–34

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Engle, R. F. (1982) 'Autoregressive Conditional Heteroscedasticity with Estimates of U.K. Inflation.' *Econometrica* 50(4), 987–1008

Engle, R. F., C. H. Hong, A. Kane, and J. Noh (1993) 'Arbitrage Valuation of Variance Forecasts with Simulated Options.' In *Advances in Futures and Options Research,* ed. D. M. Chance and R. R. Trippi (Greenwich: JAI Press)

Franq, C. and M. Roussignol and J. M. Zakoïan (2001) 'Conditional Heteroskedasticity Driven by Hidden Markov Chains.' *Journal of Time Series Analysis* 22(2), 197–220

Glosten, L. R., R. Jagannathan, and D. Runkel (1993) 'On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks.' *Journal of Finance* 48(5), 1779–1801

Gray, S. (1996) 'Modeling the Conditional Distribution of Interest Rates as a Regime-Switching Process.' *Journal of Financial Economics* 42(1), 27–62

Guidolin, M. and A. Timmermann (2005) 'Value-at-Risk and Expected Shortfall under Regime-Switching.' *UCSD mimeo*

Haas, M. and S. Mittnik and M. S. Paolella (2004a) 'A New Approach to Markov-Switching GARCH Models.' *Journal of Financial Econometrics* 2(4), 493–530

____ (2004b) 'Mixed Normal Conditional Heteroskedasticity.' *Journal of Financial Econometrics* 2(2), 211–250

Hamilton, J. D. (1988) 'Rational-Expectation Econometrc Analysis of Changes in Regime.' *Journal of Economic Dynamics and Control* 12(2-3), 385–423

____ (1989) 'A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle.' *Econometrica* 57(2), 357–384

____ (1990) 'Analysis of Time Series Subject to Changes in Regime.' *Journal of Econometrics* 45(1), 39–70

____ (1994) *Time Series Analysis* (Princeton: Princeton University Press)

Hamilton, J. D., and R. Susmel (1994) 'Autoregressive Conditional Heteroskedasticity and Changes in Regime.' *Journal of Econometrics* 64(1-2), 307–33

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Hansen, B. E. (1992) 'The Likelihood Ratio Test under Nonstandard Conditions: Testing the Markov Switching Model of GNP.' *Journal of Applied Econometrics* 7, S61–S82

____ (1994) 'Autoregressive Conditional Density Estimation.' *International Economic Review* 35(3), 705–730

____ (1996) 'Erratum: the Likelihood Ratio Test Under Nonstandard Conditions: Testing the Markov Switching Model of GNP.' *Journal of Applied Econometrics* 11(2), 195–198

Hansen, P. R. (2005) 'A Test for Superior Predictive Ability.' *forthcoming* in the *Journal of Business and Economic Statistics*

Hansen, P. R., and A. Lunde (2005) 'A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?' *forthcoming* in the *Journal of Applied Econometrics*

Harvey, D., S. Leybourne, and P. Newbold (1997) 'Testing the Equality of Prediction Mean Squared Errors.' *International Journal of Forecasting* 13(2), 281–291

Kim, C. J. (1994) 'Dynamic Linear Models with Markov-Switching.' *Journal of Econometrics* 60(1-2), 1–22

Klaassen, F. (2002) 'Improving GARCH Volatility Forecasts.' *Empirical Economics* 27(2), 363–394

Koopman, S. J., B. Jungbacker, and E. Hol (2005) 'Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements.' *Journal of Empirical Finance* 12(3), 445–475

Kupiec, P. H. (1995) 'Techniques for Verifying the Accuracy of Risk Measurement Models.' *The Journal of Derivatives* 3, 73–84

Lamoureux, C., and W. Lastrapes (1990) 'Persistence in Variance, Strucutral Change, and the GARCH Model.' *Journal of Business and Economic Statistics* 8(2), 225–234

Lin, G. (1998) 'Nesting Regime-Switching GARCH Models and Stock Market Volatility, Returns and the Business Cycle.' PhD dissertation, University of California, San Diego, San Diego

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM

Lopez, J. A. (1999) 'Methods for Evaluating Value-at-Risk Estimates.' *Federal Reserve Bank of San Francisco Economic Review* 2, 3–17

Lopez, J. A. (2001) 'Evaluating the Predictive Accuracy of Volatility Models.' *Journal of Forecasting* 20(2), 87–109

Nelson, D. B. (1991) 'Conditional Heteroskedasticity in Asset Returns: A New Approach.' *Econometrica* 59(2), 347–370

Patton, A. J. (2005) 'Volatility Forecast Evaluation and Comparison Using Imperfect Volatility Proxies.' *LSE, Financial Market Group, mimeo*

Pesaran, M. H., and A. Timmermann (1992) 'A Simple Nonparameteric Test of Predictive Performance.' *Journal of Business and Economic Statistics* 10(4), 461–465

Politis, D. N., and J. P. Romano (1994) 'The Stationary Bootstrap.' *Journal of The American Statistical Association* 89(428), 1303–1313

Sarma, M., S. Thomas, and A. Shah (2003) 'Selection of Value-at-Risk Models.' *Journal of Forecasting* 22(4), 337–358

West, K. D. (1996) 'Asymptotic Inference About Predictive Ability.' *Econometrica* 64(5), 1067–1084

West, K. D., H. J. Edison, and D. Cho (1993) 'A Utility-Based Comparison of Some Models of Exchange Rate Volatility.' *Journal of International Economics* 35(1-2), 23–45

White, H. (2000) 'A Reality Check For Data Snooping.' *Econometrica* 68(5), 1097–1126

Brought to you by | Tufts University
Authenticated
Download Date | 3/15/18 1:39 PM