Forecasting Stock Market Volatility with Regime-Switching GARCH Models

Juri Marcucci*

Department of Economics, University of California, at San Diego
9500 Gilman Drive, La Jolla CA 92093-0508, USA

This Version: March 2005

Abstract

This paper compares different GARCH models in terms of their ability to describe and forecast financial time series' volatility from one-day to one-month horizon. To take into account the excessive persistence usually found in GARCH models that implies too smooth and too high volatility forecasts, Markov Regime-Switching GARCH (MRS-GARCH) models, where the parameters are allowed to switch between a low and a high volatility regime, are analyzed. Both gaussian and fat-tailed conditional distributions for the residuals are assumed, and the degrees of freedom can be state-dependent to model possible time-varying kurtosis. The empirical analysis demonstrates that Markov Regime-Switching GARCH (MRS-GARCH) models do really outperform all standard GARCH models in forecasting volatility at shorter horizons according to a broad set of statistical loss functions. At longer horizons standard asymmetric GARCH models fare the best. Tests for equal predictive ability of the Diebold-Mariano type and for superior predictive ability, such as White's Reality Check and Hansen's SPA test, confirm these results rejecting the presence of better models than the MRS-GARCH with normal innovations. However, according to a VaR-based loss function this model fares much worse than the others at all horizons. Also we find that no model clearly outperforms the others at all horizon under this risk-management evaluation criterion.

JEL Classification: C22, C52, C53

Keywords: Markov Regime-Switching GARCH, Volatility, Forecasting, Forecast Evaluation, Fat-tailed Distributions.

*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 valuable and helpful comments that greatly helped me improve 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. E-mail: jmarcucc@weber.ucsd.edu

1 Introduction

In the last few decades a growing number of scholars has focused attention on the analysis and forecasting of volatility, due to its crucial role in financial markets. Portfolio managers, option traders and market makers all are interested in the possibility of forecasting, with a reasonable level of accuracy, this important magnitude, in order to get either higher profits or less risky positions.

So far in the literature, many models have been put forward, but those that seem to be the most successful and popular are the GARCH (Generalized Autoregressive Conditional Heterscedasticity) models by Bollerslev (1986) who generalizes the seminal idea on ARCH models by Engle (1982). Their incredible popularity stems from their ability to capture, with a very flexible structure, some of the typical stylized facts of financial time series, such as volatility clustering, that is the tendency for volatility periods of similar magnitude to cluster. Usually GARCH models can take into account the time-varying volatility phenomenon over a long period (French et al., 1987, Franses and Van Dijk, 1996) and provide very good in-sample estimates.

Furthermore, as argued by Andersen and Bollerslev (1998), GARCH models do really provide good volatility forecasts, even though it might be the case that researchers can get good in-sample fit, but very poor forecasting performances. This poor predictive power in forecasting volatility could originate both from an erroneous form of judgement, such as the practice of comparing the model forecasts to a lousy measure for the ex-post volatility, and from an incorrect choice of the statistical loss function to be minimized. For these reasons the authors put forward a new way to compare volatility forecasts, by means of the so-called 'realized volatility', calculated with intra-daily data.

One of the main goals of the present paper is to show that possible concerns about the forecasting ability of GARCH models can arise because of the usually estimated excessive persistence of individual shocks on volatility. Hamilton and Susmel (1994), for example, find that, for their 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 the GARCH volatility forecasts are sometimes too smooth and too high across periods with different levels of turbulence.

Financial returns exhibit sudden jumps that are due not only to structural breaks in the real economy, but also to changes in the operators' expectations about the future, that can originate from different information or dissimilar preferences. The real volatility is affected by millions of shocks, that anyway never persist for a long time, rendering its behavior mean-reverting. It follows that a good volatility model should entail a different way of treating shocks, in order to give better forecasts. For such reasons in the present work GARCH models are incorporated in a regime-switching framework, that allows, rather parsimoniously, to take into account the existence of two different volatility regimes, characterized by a different level of volatility. In both regimes volatility follows a GARCH-like pattern, in such a way to avoid the actual variance to depend on the entire information set, as in Gray (1996) and Klaassen (2002).

Besides, some fat-tailed distributions, such as the Student's t and the GED, are adopted. In some cases, the corresponding shape parameters can vary across different regimes, to model possible variations in the conditional kurtosis in a way that generalizes Dueker (1997)'s Regime-Switching GARCH models, where

only few parameters are state-dependent.

GARCH models typically show high volatility persistence. Lamoureux and Lastrapes (1990) attribute this persistence to the possible presence of structural breaks in the variance. They demonstrate that shifts in the unconditional variance are likely to lead to wrong estimates of the GARCH parameters in a way that implies persistence.

Cai (1994) and Hamilton and Susmel (1994) are the first to apply the seminal idea of regime-switching parameters by Hamilton (1988, 1989 and 1990) into an ARCH specification in order to account for the possible presence of structural breaks. They use an ARCH specification instead of a GARCH to avoid problems of infinite path-dependence.

The property that makes Markov Regime-Switching GARCH (heretofore MRS-GARCH) and GARCH models so different is given by their completely opposite representation of the concept of time-varying volatility. Actually, while GARCH models describe volatility as an ARMA process, thus incorporating innovations directly, the MRS-GARCH models keep the same structure for the volatility, adding the possibility of sudden jumps from the turbulent regime to the tranquil state and viceversa.

The main empirical results, using US stock market data, point out that MRS-GARCH models significantly outperform the usual GARCH models in forecasting volatility at shorter horizons according to a broad set of statistical loss functions. The significance of the performance differentials among the competing models is worked out both through Diebold-Mariano-type of tests for equal predictive ability and tests for superior predictive ability, such as the White's (2000) Reality Check test and Hansen's (2001) test for Superior Predictive Ability. Overall these tests show that the MRS-GARCH model with normal innovations does outperform all standard GARCH models in forecasting volatility at shorter horizons. At longer horizons, standard GARCH models outperform the MRS-GARCH. The tests for superior predictive ability display the predictive superiority of MRS-GARCH model with normal innovations and the results do not change if we only compare MRS-GARCH models, without including single-regime GARCH models. Since volatility is used as a key ingredient for VaR estimates a risk-management loss function is adopted to compare the forecasting performances of the competing models. According to this loss function, MRS-GARCH models seem to fare much worse than with standard statistical loss function also at shorter horizons (the only exception is the MRS-GARCH model with GED innovations). However, as in previous studies such as Brooks and Persand (2003) there is no clear answer about which model fares the best under this VaR-based loss function.

The plan of this paper is as follows. Linear and non-linear GARCH models are presented in section 2. Section 3 is devoted to a detailed description of MRS-GARCH models. The stock market data (daily and intra-daily) and the methodology are discussed in section 4, while in section 5 a digression on the various tests used to evaluate the one-day, one-week, two-week and one-month ahead volatility forecasts is presented. All the empirical results and the discussion are developed in section 6 and conclusions are sketched in section 7.

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)

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

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

$$r_t = \delta + \varepsilon_t = \delta + \eta_t \sqrt{h_t} \tag{2}$$

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta h_{t-1} \tag{3}$$

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 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

$$\log (h_t) = \alpha_0 + \alpha_1 \left| \frac{\varepsilon_{t-1}}{h_{t-1}} \right| + \xi \frac{\varepsilon_{t-1}}{h_{t-1}} + \beta_1 \log (h_{t-1}) \tag{4}$$

with no parameter constraints.

Glosten, Jagannathan and Runkle (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

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \left[1 - I_{\{\varepsilon_{t-1} > 0\}} \right] + \xi \varepsilon_{t-1}^2 I_{\{\varepsilon_{t-1} > 0\}} + \beta_1 h_{t-1} \tag{5}$$

where $I_{\{\omega\}}$ 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 a Student's t or a GED distribution. If a Student's t distribution with ν degrees of freedom is assumed, the probability density function (pdf) of ε_t takes the form

$$f(\varepsilon_t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)}(\nu-2)^{-\frac{1}{2}}(h_t)^{-\frac{1}{2}}\left[1+\frac{\varepsilon_t^2}{h_t(\nu-2)}\right]^{-\frac{(\nu+1)}{2}} \tag{6}$$

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. With a GED distribution instead, the pdf of the innovations becomes

$$f(\varepsilon_t) = \frac{\nu \exp\left[-\left(\frac{1}{2}\right)\left|\frac{\varepsilon_t}{\lambda h_t^{1/2}}\right|^{\nu}\right]}{h_t^{1/2}\lambda 2^{\left(1+\frac{1}{\nu}\right)}\Gamma\left(\frac{1}{\nu}\right)} \tag{7}$$

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 $0 < \nu \leq \infty$ and indicating how thick the tails of the distribution are, compared to the normal. 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 of the model to switch across different regimes (or *states* of the world) according to a Markov process, which is governed by a state variable, denoted s_t. The logic behind this kind of modeling 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 such observation. The state variable is assumed to evolve according to a first-order Markov chain, with transition probability

$$\Pr(s_t = j | s_{t-1} = i) = p_{ij} \tag{8}$$

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

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{bmatrix} = \begin{bmatrix} p & (1-q) \\ (1-p) & q \end{bmatrix} \tag{9}$$

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 $\pi_1 = (1-p)/(2-p-q)$.

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

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

$$r_t|\zeta_{t-1} \sim \begin{cases} f\left(\theta_t^{(1)}\right) & \text{w. p. } p_{1,t} \\ f\left(\theta_t^{(2)}\right) & \text{w. p. } (1-p_{1,t}) \end{cases} \tag{10}$$

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, $p_{1,t} = \Pr[s_t = 1|\zeta_{t-1}]$ is the ex ante probability and ζ_{t-1} denotes the information set at time $t-1$, that is the σ-algebra induced by all the variables that are observed at $t-1$. More specifically, the vector of time-varying parameters can be decomposed into three components

$$\theta_t^{(i)} = \left(\mu_t^{(i)}, h_t^{(i)}, \nu_t^{(i)}\right) \tag{11}$$

where $\mu_t^{(i)} \equiv E(r_t|\zeta_{t-1})$ is the conditional mean (or location parameter), $h_t^{(i)} \equiv Var(r_t|\zeta_{t-1})$ is the conditional variance (or scale parameter), and $\nu_t^{(i)}$ is the shape parameter of the conditional distribution.[2] 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

$$r_t = \mu_t^{(i)} + \varepsilon_t = \delta^{(i)} + \varepsilon_t \tag{12}$$

where $i = 1, 2$ and $\varepsilon_t = \eta_t\sqrt{h_t}$ and η_t is a zero mean, unit variance process. The main reason for this choice is given by our main 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[3] $h_t^{(i)} = V[\varepsilon_t|\widetilde{s}_t, \zeta_{t-1}]$. For this conditional variance the following GARCH(1,1)-like expression is assumed

$$h_t^{(i)} = \alpha_0^{(i)} + \alpha_1^{(i)}\varepsilon_{t-1}^2 + \beta_1^{(i)}h_{t-1} \tag{13}$$

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 be a function of all past states.

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

[3] Here we are using Klaassen's (2002) model simplifying his notation.

Cai (1994) and Hamilton and Susmel (1994) are the first to point out this difficulty by combining the regime-switching approach with ARCH models only, thus eliminating the GARCH term in (13). 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 (13) 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

$$\begin{aligned} h_{t-1} = E_{t-2}\{h_{t-1}^{(j)}\} = p_{1,t-1}\left[\left(\mu_{t-1}^{(1)}\right)^2 + h_{t-1}^{(1)}\right] + (1-p_{1,t-1})\left[\left(\mu_{t-1}^{(2)}\right)^2 + h_{t-1}^{(2)}\right] \\ - \left[p_{1,t-1}\mu_{t-1}^{(1)} + (1-p_{1,t-1})\,\mu_{t-1}^{(2)}\right]^2 \end{aligned} \tag{14}$$

where $j = 1, 2$. The main drawback of this specification is its inconvenience in terms of volatility forecasting, because 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

$$\begin{aligned} \frac{\sigma_t^\nu - 1}{\nu} =& \omega_{s_{t_1}} + \alpha_{s_{t_2}}(L)_p\,\widetilde{\sigma}_{t-1}^\nu\,|f(\varepsilon_{t-1})|^w - \lambda_{s_{t_2}}\widetilde{\sigma}_{t-1}^\nu\,|f(\varepsilon_{t-1})|^w\,\frac{\varepsilon_{t-1}}{|\varepsilon_{t-1}|} + \\ & \beta_{s_{t_3}}(L)_q\left[\frac{\widetilde{\sigma}_{t-1}^\nu - 1}{\nu}\right] \end{aligned} \tag{15}$$

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) 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 by also taking into account the current one, Klaassen (2002) adopts the following expression for the conditional variance

$$h_t^{(i)} = \alpha_0^{(i)} + \alpha_1^{(i)}\varepsilon_{t-1}^2 + \beta_1^{(i)} E_{t-1}\{h_{t-1}^{(i)}|s_t\} \tag{16}$$

where the expectation is computed as

$$E_{t-1}\{h_{t-1}^{(i)}|s_t\} = \widetilde{p}_{ii,t-1}\left[\left(\mu_{t-1}^{(i)}\right)^2 + h_{t-1}^{(i)}\right] + \widetilde{p}_{ji,t-1}\left[\left(\mu_{t-1}^{(j)}\right)^2 + h_{t-1}^{(j)}\right] - \left[\widetilde{p}_{ii,t-1}\mu_{t-1}^{(i)} + \widetilde{p}_{ji,t-1}\mu_{t-1}^{(j)}\right]^2 \tag{17}$$

and the probabilities are calculated as

$$\widetilde{p}_{ji,t} = \Pr\left(s_t = j | s_{t+1} = i, \zeta_{t-1}\right) = \frac{p_{ji} \Pr\left(s_t = j | \zeta_{t-1}\right)}{\Pr\left(s_{t+1} = i | \zeta_{t-1}\right)} = \frac{p_{ji} p_{j,t}}{p_{i,t+1}} \tag{18}$$

with $i, j = 1, 2$.

Klaassen's (2002) regime-switching GARCH has two main advantages over the other models. Within the model, it allows higher flexibility in capturing the persistence of shocks to volatility.[4] 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 h-step ahead volatility forecast at time $T-1$ can be calculated as follows

$$\widehat{h}_{T,T+h} = \sum_{\tau=1}^{h} \widehat{h}_{T,T+\tau} = \sum_{\tau=1}^{h} \sum_{i=1}^{2} \Pr\left(s_\tau = i | \zeta_{T-1}\right) \widehat{h}^{(i)}_{T,T+\tau} \tag{19}$$

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

$$\widehat{h}^{(i)}_{T,T+\tau} = \alpha_0^{(i)} + \left(\alpha_1^{(i)} + \beta_1^{(i)}\right) E_T\{h^{(i)}_{T,T+\tau-1} | s_{T+\tau}\} \tag{20}$$

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. Each regime volatility forecast is obtained with a GARCH-like formula where the expectation of the previous period volatility is determined by weighting the previous regime volatilities with the probabilities in (18). In general, to compute the volatility forecasts the filter probability at τ periods ahead $\Pr\left(s_{t+\tau} = i | \zeta_t\right) = p_{i,t+\tau} = P^\tau p_{i,t}$ is needed.

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

$$\begin{aligned} p_{1,t} = \Pr\left[s_t = 1 | \zeta_{t-1}\right] = (1-q) & \left[\frac{f\left(r_{t-1} | s_{t-1} = 2\right)\left(1 - p_{1,t-1}\right)}{f\left(r_{t-1} | s_{t-1} = 1\right) p_{1,t-1} + f\left(r_{t-1} | s_{t-1} = 2\right)\left(1 - p_{1,t-1}\right)}\right] \\ + p & \left[\frac{f\left(r_{t-1} | s_{t-1} = 1\right) p_{1,t-1}}{f\left(r_{t-1} | s_{t-1} = 1\right) p_{1,t-1} + f\left(r_{t-1} | s_{t-1} = 2\right)\left(1 - p_{1,t-1}\right)}\right] \end{aligned} \tag{21}$$

where p and q are the transition probabilities in (9) and $f(\cdot)$ is the likelihood given in (10).

Thus, the log-likelihood function can be written as

[4] 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.

$$\ell = \sum_{t=-R+w+1}^{T+w} \log \left[p_{1,t} f\left(r_t | s_t = 1\right) + \left(1 - p_{1,t}\right) f\left(r_t | s_t = 2\right) \right] \tag{22}$$

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 October 15, 2003 for a total of 4095 observations all obtained from Datastream. The sample is divided in two parts. The first 3985 observations (from January 1, 1988 to September 29, 2001) are used as the in-sample for estimation purposes, while the remaining 511 observations (from October 1, 2001 to October 15, 2003) are taken as the out-of-sample for forecast evaluation purposes.

Table 1 shows some descriptive statistics of the S&P100 rate 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 empirical distribution of the returns is longer than the upper tail, that is negative returns are more likely to be far below the mean than their counterparts.

[INSERT TABLE 1 HERE]

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. Both the conditional mean and the conditional variances are estimated jointly by maximizing the log-likelihood function which is computed as the logarithm of the product of the conditional densities of the prediction errors as shown in (22).

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[5].

The "true volatility" would be needed to evaluate the forecasting performances of 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 a very noisy estimate of the unobserved volatility. As a matter of fact it can lead to wrong and

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

unfair assessments about the real ability of various GARCH models in forecasting volatility. As highlighted in Andersen and Bollerslev (1998) one possibility to avoid such bad conclusions about the relatively poor out-of-sample performances of GARCH models 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 either of few minutes or of few hours.

In this paper we adopt three different measures of the acutal volatility denoted $\widehat{\sigma}^2_{t+1|t}$. The first one is the realized volatility computed as the sum of 5-minute returns over each day. Intra-daily returns on the S&P100 are obtained from www.disktrading.com. These data are also used by Hol and Koopman (2005). To calculate the volatility at h-step ahead we sum the daily realized volatility over the h 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 1-day. The third measure is given by the squared return of the forecasting horizon. Thus, if for example we are forecasting volatility at 1-week, we use the square of the log difference of closing price at time t and $t+5$.

We denote the h-step ahead volatility forecast as $\widehat{h}_{t+h|t}$ which is computed as the aggregated sum of the forecasts for the next h steps made at time t, i.e. $\widehat{h}_{t+h|t} = \sum_{j=1}^{h} \widehat{h}_{t+j|t}$. We compute volatility forecasts at 1-day, 5-, 10- and 22-days by aggregating the volatility forecast over the next 1, 5, 10 and 22 days. Actually practitioners and risk managers are not interested in the multi-step ahead one-day volatility forecasts, such as the volatility 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 given by 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, Engle and Nelson (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, Egle, Hong, Kane and Noh (1993) and West, Edison, and Cho (1993) suggest profit-based and utility-based 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 statistical loss function, the Mean Squared Error (MSE).

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:

$$MSE_1 = n^{-1} \sum_{t=1}^{n} \left(\widehat{\sigma}_{t+1} - \widehat{h}^{1/2}_{t+1|t} \right)^2 \tag{23}$$

$$MSE_2 = n^{-1}\sum_{t=1}^{n}\left(\widehat{\sigma}^2_{t+1} - \widehat{h}_{t+1|t}\right)^2 \tag{24}$$

$$QLIKE = n^{-1}\sum_{t=1}^{n}\left(\log \widehat{h}_{t+1|t} + \widehat{\sigma}^2_{t+1}\widehat{h}^{-1}_{t+1|t}\right) \tag{25}$$

$$R2LOG = n^{-1}\sum_{t=1}^{n}\left[\log\left(\widehat{\sigma}^2_{t+1}\widehat{h}^{-1}_{t+1|t}\right)\right]^2 \tag{26}$$

$$MAD_1 = n^{-1}\sum_{t=1}^{n}\left|\widehat{\sigma}_{t+1} - \widehat{h}^{1/2}_{t+1|t}\right| \tag{27}$$

$$MAD_2 = n^{-1}\sum_{t=1}^{n}\left|\widehat{\sigma}^2_{t+1} - \widehat{h}_{t+1|t}\right| \tag{28}$$

$$HMSE = T^{-1}\sum_{t=1}^{T}\left(\widehat{\sigma}^2_{t+1}\widehat{h}^{-1}_{t+1|t} - 1\right)^2 \tag{29}$$

The criteria in (23) and (24) are the typical mean squared error metrics. The criteria in (24) and (26) are exactly equivalent to using the R^2 metric in the Mincer-Zarnowitz regressions of $\widehat{\sigma}^2_{t+1}$ on a constant and $\widehat{h}_{t+1|t}$ and of $\log\left(\widehat{\sigma}^2_{t+1}\right)$ on a constant and $\log\left(\widehat{h}_{t+1|t}\right)$, respectively, provided that the forecasts are unbiased. Moreover the $R2LOG$ loss function has the particular feature of penalizing volatility forecasts asymmetrically in low volatility and high volatility periods, as pointed out by Pagan and Schwert (1990) who put forward (26), calling it logarithmic loss function. The loss function in (25) corresponds to the loss implied by a gaussian likelihood and is suggested by Bollerslev, Engle and Nelson (1994). The Mean Absolute Deviation (MAD) criteria in (27) and (28) 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 (29).

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[6] 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, Hong, Kane and Noh, 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 $\overline{\sigma}_{t+j}$ be the proxy for the actual volatility after subtracting its non-zero mean and let $\overline{h}_{t+j|t+j-1}$ be the demeaned volatility forecasts[7]. The SR is simply the fraction of the volatility forecasts that have the

[6] We talk about direction of change, because volatility is always positive.

[7] The author acknowledges that because of the Jensen's inequality, $E(X^2) \geq [E(X)]^2$, such a procedure can give results

same direction of change as the corresponding realizations and is given by

$$SR = m^{-1} \sum_{j=1}^{m} I_{\{\overline{\sigma}_{t+j} \overline{h}_{t+j|t+j-1}\}>0} \tag{30}$$

where $I_{\{g>0\}}$ is the indicator function, that is $I_{\{g>0\}} = 1$ 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

$$DA = \frac{(SR - SRI)}{\sqrt{Var(SR) - Var(SRI)}} \tag{31}$$

where $SRI = P\widehat{P} + (1-P)(1-\widehat{P})$ and P represents the fraction of times that $\overline{\sigma}_{t+j} > 0$, while $\widehat{P}$ is the proportion of demeaned volatility forecasts that are positive. $Var(SR)$ and $Var(SRI)$ are the corresponding variances. The Directional Accuracy 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 test or with a joint test. When one compares the predictive ability of pairs of competing models, the usual test employed is the Diebold-Mariano 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 one can only test two different models and decide which one is better. The test that we can adopt in this case is the Reality Check of White (2000) or the Superior Predictive Ability test of Hansen (2001).

Diebold and Mariano (DM) (1995) propose a test of equal predictive ability (EPA heretofore) 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) showed 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 $k = 1, \ldots, q$[8].

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.

[8]Because for optimal h-step ahead forecasts the sequence of forecast errors follows a MA of order $h-1$. q has been chosen so

The DM test statistic for testing the null hypothesis of equal forecast accuracy is then given by $DM = \overline{d}/\sqrt{\widehat{V}\left(\overline{d}\right)} \sim N\left(0,1\right)$, i.e. under the null hypothesis of equal forecast accuracy the DM test statistic has a standard normal distribution asymptotically. Harvey, Leybourne and Newbold (1997) argued 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) test, where DM is multiplied by $\sqrt{n^{-1}\left[n+1-2h+n^{-1}h(h-1)\right]}$, where h is the forecast horizon and n is the length of the evaluation period[9]

Instead of testing for EPA, as in Diebold and Mariano (1995) or in West (1996), the Reality Check (RC) for data snooping of White (2000) is a test for superior predictive ability [10] (SPA).

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 forecasting 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 $k=1,\ldots,l$ outperform 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[11] 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

$$f_{k,t} = L_{t,0} - L_{t,k} \qquad k=1,\ldots,l \quad t=1,\ldots,n \tag{32}$$

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 are better than the benchmark (that is no predictive superiority over the benchmark itself) can be equivalently formulated as

$$H_0 : \ \mu_{\max} \equiv \max_{k=1,\ldots,l} \mu_k \leq 0 \tag{33}$$

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

By the law of large numbers one 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

---

that $q = \lfloor 4 * (n/100)^{2/9} \rfloor$.

[9] Harvey, Leybourne and Newbold (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.

[10] 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.

[11] 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 (24).

$$T_n \equiv \max_{k=1,\ldots,l} n^{1/2}\overline{f}_{k,n} \tag{34}$$

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 (2001) emphasizes that the Reality Check test applies a supremum over the non-standardized performances T_n and, more dangerously, a conservative asymptotic distribution that makes the RC very sensitive to the inclusion of poor models. Hansen (2001) 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. Hansen (2001) applies a supremum over the standardized performances and tests the null hypothesis

$$H_0 : \mu^s_{\max} \equiv \max_{k=1,\ldots,l} \frac{\mu_k}{\sqrt{var(n^{1/2}\overline{f}_{k,n})}} \leq 0 \tag{35}$$

using the statistic

$$T^s_n = \max_k \frac{n^{1/2}\overline{f}_{k,t}}{\sqrt{\widehat{var}(n^{1/2}\overline{f}_{k,n})}} \tag{36}$$

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 (2001) suggests additional refinements to the RC test and some modifications of the asymptotic distribution that result in tests less sensitive to the inclusion of poor models and with a better power. He argues as well that the p-values of the RC are generally inconsistent (that is too large) and the test can be asymptotically biased. To overcome these drawbacks, Hansen (2001) shows that it is possible to derive a consistent estimate of the p-value together with an upper and a lower bound. Such a test is called Superior Predictive Ability (SPA) test and it includes the RC as a special case. The upper bound (SPA_u) is the p-value of a conservative test (that is, it has the same asymptotic distribution as the RC test) where it is implicitly assumed that all the competing models $(k = 1, \ldots, l)$ are as good as the benchmark in terms of expected loss. Hence, the upper bound p-value coincides with the RC test p-value. The lower bound (SPA_l) is the p-value of the liberal test where the null hypothesis assumes that the models with worse performance than the benchmark are poor models in the limit. With the SPA test it is possible 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 Reality Check test) is quite sensitive to the inclusion of poor and irrelevant models in the comparison, while the consistent (SPA_c) and the liberal test are not[12].

[12] For a detailed description of how to implement the RC and SPA test, see White (2000) and Hansen (2001).

5.3 Risk Management Loss Functions

Since one of the typical use of volatility forecasts is as an input to financial risk management, we also employ a risk management loss function, which is based upon the calculation of the Value at Risk (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) suggest to use VaR-based loss functions and also Sarma, Thomas and Shah (2002).

The VaR at time t of model i at α% significance level is calculated as follows

$$VaR_t^i\left[n,\alpha\right]=\mu_{t+n}^i+\Phi\left(\alpha\right)\sqrt{h_{t+n}^i} \tag{37}$$

where $\Phi()$ is a cumulative distribution function, n is the investment horizon ($n=1,5,10,20$ days), α =1% or 5%, μ_{t+n}^i is the conditional mean at $t+n$ and h_{t+n}^i is the volatility forecast at $t+n$ of model i.

We thus employ three methods to determine the adequacy of the volatility forecasts used as an input for the VaR.

The TUFF test is based on the number of observations before the first exception. The relevant null is, once again, $H_0:\alpha=\alpha_0$ and the corresponding LR test is

$$LR_{TUFF}(\widetilde{T},\widehat{\alpha})=-2\log\{\widehat{\alpha}\left(1-\widehat{\alpha}\right)^{\widetilde{T}-1}\}+2\log\{\frac{1}{\widetilde{T}}\left(1-\frac{1}{\widetilde{T}}\right)^{\widetilde{T}-1}\} \tag{38}$$

where $\widetilde{T}$ denotes the number of observations before the first exception. LR_{TUFF} is also asymptotically distributed as $\chi^2(1)$. Kupiec (1995) notices that this test has limited power to distinguish among alternative hypotheses because all observations after the first failure are ignored, resulting in a test which is over-sized.

A correctly specified VaR model should generate the pre-specified failure rate conditionally at every point in time. This property is known as "conditional coverage" of the VaR. Christoffersen (1998) develops a framework for interval forecast evaluation. The VaR is interpreted as a forecast that the portfolio return will lie in $(-\infty,VaR_t)$ with a pre-specified probability p. Christoffersen emphasizes the importance of testing for conditional coverage due to the well known stylized fact in financial time series of volatility clustering. Good interval forecasts should be narrow in tranquil periods and wide in volatile times, so that observations falling outside a forecasted interval should be spread over the entire sample and not concentrated in clusters. A poor interval forecast may produce correct unconditional coverage, but it may fail to account for higher-order dynamics. In such case the symptom that could be observed is a clustering of failures.

Consider a sequence of one-period-ahead VaR forecasts $\{v_{t|t-1}\}_{t=1}^T$, estimated at a significance level $1-p$. These forecasts are intended to be one-sided interval forecasts $(-\infty,v_{t|t-1}]$ with coverage probability p. Given the realizations of the return series r_t and the ex-ante VaR forecasts, the following indicator variable can be calculated

$$I_t=\begin{cases}1, & r_t<v_t\\ 0, & \text{otherwise}\end{cases}$$

The stochastic process $\{I_t\}$ is called 'failure process'. The VaR forecasts are said to be efficient if they display "correct conditional coverage", i.e. if $E\left[I_{t|t-1}\right] = p, \forall t$ or, equivalently, if $\{I_t\}$ is *iid* with mean p.

Christoffersen (1998) develops a three step testing procedure to test for correct conditional coverage: (i) a test for correct unconditional coverage, (ii) a test for independence and (iii) a test for correct conditional coverage.

In the test for correct unconditional coverage the null hypothesis of the failure probability p is tested against the alternative hypothesis that the failure probability is different from p, under the assumption of and independently distributed failure process. In the test for independence, the hypothesis of an independently distributed failure process is tested against the alternative hypothesis of a first order Markov failure process. Finally, the test of correct conditional coverage is done by testing the null hypothesis of an independent failure process with failure probability p against the alternative of a first order Markov failure process.

All the three tests are carried out in the likelihood ratio (LR) framework. The likelihood ratio for each is as follows:

1. LR statistic for the test of unconditional coverage:

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

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

2. LR statistic for the test of independence:

$$LR_{ind} = -2\log\frac{\left(1-\widehat{\pi}_2\right)^{(n_{00}+n_{10})}\left(1-\widehat{\pi}_2\right)^{(n_{01}+n_{11})}}{\left(1-\widehat{\pi}_{01}\right)^{n_{00}}\widehat{\pi}_{01}^{n_{01}}\left(1-\widehat{\pi}_{11}\right)^{n_{10}}\widehat{\pi}_{11}^{n_{11}}} \sim \chi^2_{(1)}$$

   where n_{ij} is the number of i values followed by a j value in the I_t series $(i,j = 0,1)$, $\pi_{ij} = \Pr\{I_t = i|I_{t-1} = j\}$ $(i,j = 0,1)$, $\widehat{\pi}_{01} = n_{01}/(n_{00}+n_{01})$, $\widehat{\pi}_{11} = n_{11}/(n_{10}+n_{11})$, $\widehat{\pi}_2 = (n_{01}+n_{11})\,/\,(n_{00}+n_{01}+n_{10}+n_{11})$.

3. LR statistic for the test of correct conditional coverage:

$$LR_{cc} = -2\log\frac{\left(1-p\right)^{n_0}p^{n_1}}{\left(1-\widehat{\pi}_{01}\right)^{n_{00}}\widehat{\pi}_{01}^{n_{01}}\left(1-\widehat{\pi}_{11}\right)^{n_{10}}\widehat{\pi}_{11}^{n_{11}}} \sim \chi^2_{(2)}$$

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

In the VaR-based forecast evaluation we will use all these tests. For the interpretation of the results, models are ranked in this way. As in Brooks and Persand (2003) we assume that any model which has a percentage failures in the rolling hold-out sample which is greater than the nominal threshold should

be rejected as inadequate. Thus, the lowest ranking models (the worst) are those which have the highest percentage of failures greater than the nominal value. When these models have been exhausted, we assume that any model that generates a number of failures which is far less than the expected number should be less desirable than those models which present a number of failures closer to the nominal level. Hence, the best models under this loss function are those which generate a coverage rate which is less than the nominal one, but as close as possible to it. Then we check for correct unconditional and conditional coverage. If a model can also pass these tests is considered adequate for risk-management purposes.

6 Empirical Results and Discussion

The whole sample consists of the S&P100 closing prices from January 1, 1988 to October 15, 2003, for a total of 4096 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 3085 observations. In the present section we present the empirical estimates of single-regime GARCH and MRS-GARCH models, together with the in-sample statistics and the out-of-sample forecast evaluation.

6.1 Single-regime GARCH

The parameter estimates for 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 for 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 for 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.

[TABLE 2 HERE]

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 that for this distribution is given by $\left(\Gamma\left(1/\nu\right)\Gamma\left(5/\nu\right)\right)/\left(\Gamma\left(1/\nu\right)\right)^2)$ 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.

6.2 MRS-GARCH

The parameter estimates for 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. The estimates confirm the existence of two states: the first regime is characterized by a low volatility and almost nil persistence of the shocks in the conditional volatility, whereas the second one reveals high volatility and a higher persistence. The transition probabilities are all significant but in the normal case that is rather far away from unity, showing that almost all regimes are particularly persistent.

[TABLE 3 HERE]

Table 3 also reports the unconditional probabilities and the expected durations for each MRS-GARCH model. The unconditional probability π_1 of being in the first regime, which is characterized overall by a lower volatility than in the second one, ranges between 2% for the model with gaussian innovations and 61% for the model with GED innovations. The expected duration for this low-volatility regime ranges between 53 and 8771 trading days. On the other hand, the unconditional probability of being in the high-volatility regime (the second one) ranges between 39% for the GED model and 98% for the model with Normal innovations. The expected duration of the high-volatility state is roughly between one day and 7519 days. 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 ν parameter is below the threshold value of 2, showing that the distribution has thicker tails than the normal. The conditional kurtosis for 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 are allowed to switch, implying a time-varying kurtosis as in Hansen (1994) and Dueker (1997). There is a main difference between those papers and the present work. While Hansen suggests a model in which the Student's t degrees-of-freedom parameter is allowed to vary 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 the degrees-of-freedom parameter is allowed to switch across regimes together with all the other parameters. Since both regimes show an estimated number of degrees of freedom greater than 4, we can argue that in both regimes 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[13]. Since the main focus is on the predictive ability, we only present some statistics

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

in Table 4, without doing any formal test.

[TABLE 4 HERE]

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 are allowed to switch.

The Akaike Information Criterion (AIC) and the Schwarz Criterion (BIC) both indicate that the best model among the constant-parameter GARCH and overall is the EGARCH with GED errors, while among the MRS-GARCH models is the MRS-GARCH(1,1)-t2 that fits the best. Another property of MRS-GARCH models that emerges from Table 4 is the high persistence of the shocks in 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 in the present work 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 competing models.

Since most models only represent simple approximations of the true data generating process, often having 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 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 ten- and twenty-two-step ahead volatility forecasts. For both tables the true volatility is given by the realized volatility.

[TABLES 5 AND 6 HERE]

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-step ahead, the best model is the MRS-GARCH-N and the second best model is the GJR-N. At five-step ahead, the best model is again the MRS-GARCH-N, while the second best is the EGARCH-N.

At two-week horizon, the best model is the EGARCH-N and the MRS-GARCH-N is just the best model among the MRS-GARCH. At one-month horizon, the best model is the EGARCH-GED, while the MRS-GARCH-N is only the best among the MRS-GARCH[14]

From 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 one-day horizon and the EGARCH-N at two-week horizon.

Table 7 report the DM test when the benchmark is the best model for one-day horizon (MRS-GARCH-N) which is 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. From the table 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 loss 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.

[TABLE 7 HERE]

Table 8, instead, presents the DM test when the benchmark model is the best at 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.

[TABLE 8 HERE]

The results for all other models and forecast horizons[15] 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 outperforms the benchmark at horizon longer than one-day. When the benchmark is the EGARCH model,

[14]When the proxy for the volatility is the d-day squared return, at all one-day, one-, two-, four-week horizon 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 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 one-week horizon the best model is the GJR-t whereas the best among the MRS-GARCH (MRS-GARCH-t2) is the eighth. At two-week horizon, the best model is the GJR-t, while the best among the MRS-GARCH (MRS-GARCH-GED) is sixth). At one-month horizon the best model is the GJR-t and the best among the MRS-GARCH (MRS-GARCH-GED) is fourth. All the corresponding tables are available upon request from the author.

[15]For the 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 $\sqrt{n^{-1}\left[n+1-2h+n^{-1}h(h-1)\right]}$ is .95, .98, .99, .99 for one-, five-, ten- and twenty-two step ahead horizon respectively. These results are also available upon request.

it outperforms 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 benchmark, 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 outperform 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 other models till the one-week horizon, whereas it is beaten by the EGARCH-N and EGARCH-GED at longer horizons. However, it fares the best if compared to the other MRS-GARCH models. 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 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 also other standard GARCH models, such as the EGARCH-N, EGARCH-GED or GJR-N, tend to be superior. Another striking feature of all this pairwise analysis 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 Reality Check test for superior predictive ability for each model against all the others at 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 (2001) consistent and lower p-values, respectively[16].

[TABLE 9 HERE]

The p-values reported in Table 9 for the RC and SPA tests distinctly show how all the 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, among those considered, 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 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 (2001) obtain similar results with stock market data, finding that the GARCH(1,1) specification is not the best model (in term 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 and the comparison is still carried out against all the other models. It is evident that the MRS-GARCH-N model significantly outperforms all the other models at the usual significance level

[16] Such p-values are calculated adopting the stationary bootstrap by Politis and Romano (1994) as in White (2000) and in Hansen and Lunde (2001). 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 $B = (1000, 3000)$ and a different set of values for q (0.10, 0.20 and 0.33). The results do not change considerably. Therefore we choose to report the table with $B = 3000$ and $q = 0.33$. The other tables are available upon request.

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 model for which we cannot reject the null of SPA. Similar results are obtained for all the other forecast horizons and different block length. In general, MRS-GARCH-N is always the best model according to all 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 HERE]

Table 10 reports the same RC and SPA test p-values when the comparison is done only among the MRS-GARCH models' two-week ahead volatility forecasts. This table can thus help us understand the possible implications of including poor models for these tests. The results are quite 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 other MRS-GARCH for every loss but HMSE and the MRS-GARCH-GED also fares the best according to some loss functions.

These and the previous results must also be compared to a VaR-based evaluation criterion. Since one of the main purpose of volatility forecasting is to have an input for successive VaR estimation, it is necessary to see how competing models do fare in terms of a risk-management loss function. 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 misleading. The authors propose to adopt alternative economic loss functions. 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 unconditional and conditional coverage of the corresponding VaR estimates. Sarma, Thomas and Shah (2003) go even further by adopting a second stage selection criterion of their VaR models using subjective loss functions that should incorporate the risk manager's preferences. This loss functions take into account the magnitude of the failure in the VaR forecast, penalizing more the bigger ones.

Table 11 reports the risk-management out-of-sample evaluation of our competing GARCH models for the 1-day, 1-week, 2-week and 1-month horizons.

Five statistics are presented for each forecast horizon: the TUFF, the proportion of failures (PF), the test of correct unconditional coverage (LR_{PF}) which checks if PF is significantly higher than the nominal rate, the LR_{ind} which tests independence and the LR_{cc} which tests the correct conditional coverage.

The rank for each forecast horizon gives the order according to the percentage PF. Models which present a PF greater than the coverage probability (5 and 1%) are judged as inadequate. The theoretical TUFF at 5 and 1% should be 20 and 100, respectively. We can thus see that only for 99% VaR at one-day ahead

there are values under the theoretical ones except for the MRS-GARCH-t2 and MRS-GARCH-t. At all the other forecast horizons the TUFF is greater than 100. In addition, if the objective is to cover either the 99% or the 95% of future losses, then many models seem inadequate, especially at the shortest and longest forecast horizons. The last three LR tests reject almost all models at longer 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 other MRS-GARCH models fare better according to this loss function, even though only the MRS-GARCH-GED model is not rejected by all three LR tests at 1-day horizon. At this horizon, however, other standard GARCH models also fail to reject the three LR tests showing good out-of-sample performances under the risk-management loss function. Nevertheless, it is not really clear which model among these fares the best. At longer horizons, no model can really pass all tests. Therefore, a few models seem to provide reasonable and accurate VaR estimates at 1-day horizon, with a coverage rate close to the nominal level. Actually, there is not a uniformly most accurate model according to the risk-management loss functions. This result is not new, since also Brooks and Persand (2003) find a no clear answer for most of the series they examine. Somehow, our results confirm Dacco and Satchell's (1999) arguments that the choice of the correct loss function is fundamental for the accuracy of volatility forecasts from non-linear models.

[FIGURES 2 and 3 HERE]

Figures 2 and 3 depict the excessive losses of 95% VaR and 99% VaR from GRJ-t and MRS-GARCH-t2 models. It is not clear the differences in the performance of the two models. The MRS-GARCH-t2 model seems worse than the GJR-t to capture quickly the changes in volatility of the returns.

Figure 1 illustrates the volatility forecasts at 1-day, 1-week, 2-week and 1-month horizons from the best models according the statistical out-of-sample evaluation when the proxy for volatility is the 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 quite evident at the shorter horizons that standard GARCH model's volatility forecasts tend to have higher spikes than those of the MRS-GARCH, while the reverse is true at longer forecast horizons. Thus the model which fares the best usually gives much smoother forecasts than the other.

[FIGURE 1 HERE]

In sum, no model seems to outperform all the others in forecasting volatility according to the different out-of-sample evaluation criteria adopted. Therefore accounting for regime shifts in all the parameters of the first two moments of the conditional distribution of US stock returns, together with the inclusion of GARCH effects and fat-tailed-ness gives a better in-sample fit, and outstanding out-of-sample results according to the usual statistical loss functions. However, when a more realistic loss function is used, such as a risk-management loss function, the results confirm the 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 an economic loss function. 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 non-linear models. Therefore, the practical relevance of regime-switching models in predicting the volatility turns out to be dubious. Further research is needed to evaluate these highly non-linear models according to loss functions that should capture what is really relevant for the final use of the volatility forecast.

7 Conclusions

In this paper we compare a set of standard GARCH models and Markov Regime-Switching GARCH in terms of their ability to forecast US stock market volatility. The standard GARCH models considered are 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 is allowed to switch between two regimes, one characterized by a lower volatility than the other. In addition, all models are estimated assuming both gaussian innovations and fat-tailed distributions, such as the Student's t and the GED. Further, to model time-varying conditional kurtosis, the degrees-of-freedom parameter in the Student's t distribution is allowed to switch across the two different regimes in a completely different setting than the one considered by Hansen (1994) or 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 S&P100. Such out-of-sample comparison is carried out by comparing the one-day, one-week, two-week and one-month ahead volatility forecasts.

The proxy for the true volatility is given by the realized volatility calculated by aggregating five-minute returns. The forecasting performances of each model are measured using both statistical and VaR-based loss functions.

Overall, the empirical results show that MRS-GARCH models significantly outperform standard GARCH models in forecasting volatility at shorter horizons according to a broad set of statistical loss functions. This strong conclusion is drawn considering whether the difference in the performances is significant or not. In order to test that, we apply the pairwise test for equal predictive ability of the Diebold-Mariano type. We also apply the more general Reality Check for superior predictive ability of White (2000) and the test for Superior Predictive Ability of Hansen (2001). According to these tests the MRS-GARCH-N model outperforms all other competing models.

Since volatility forecasting is mainly used as an input for successive VaR estimation, we also evaluate the competing models out-of-sample according to a VaR-based loss function. A few models seem to provide reasonable and accurate VaR estimates at 1-day horizon, with a coverage rate close to the nominal level, but there is not a uniformly most accurate model according to the risk-management loss functions. This result is not new, since also Brooks and Persand (2003) find a no clear answer for most of the series they examine. Somehow, our results also confirm Dacco and Satchell's (1999) arguments that the choice of the correct loss function is fundamental for the accuracy of volatility forecasts from non-linear models.

In sum, no model seems to outperform all the others in forecasting volatility according to the different

out-of-sample evaluation criteria adopted. Therefore accounting for regime shifts in all the parameters of the first two moments of the conditional distribution of US stock returns, together with the inclusion of GARCH effects and fat-tailed-ness gives a better in-sample fit, and outstanding out-of-sample results according to the usual statistical loss functions. However, when a more realistic loss function is used, such as a risk-management loss function, the results confirm the 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 an economic loss function. 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 non-linear models. Therefore, the practical relevance of regime-switching models in predicting the volatility turns out to be dubious. Further research is needed to evaluate these highly non-linear models according to loss functions that should capture what is really relevant for the final user of the volatility forecast.

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

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

Bollerslev, T., and E. Ghysels (1996) 'Periodic Autoregressive Conditional Heteroskedasticity.' *Journal of Business and Economic Statistics* 14, 139–157

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, 1–22

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

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

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

Franses, P. H., and R. Van Dijk (1996) 'Forecasting Stock Market Volatility Using (Non-Linear) GARCH Models.' *Journal of Forecasting* 15(3), 229–35

French, K. R., G. W. Schwert, and R. F. Stambaugh (1987) 'Expected Stock Returns and Volatility.' *Journal of Financial Economics* 19, 3–30

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

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

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

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

____ (1990) ‘Analysis of Time Series Subject to Changes in Regime.’ *Journal of Econometrics* 45, 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, 307–33

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, 195–198

Hansen, P. R. (2001) ‘An Unbiased and Powerful Test of Superior Predictive Ability.’ mimeo, Brown University

Hansen, P. R., and A. Lunde (2001) ‘A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?’ mimeo, Brown University

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

Hol, E., and S. J. Koopman (2005) ‘Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements.’ *Journal of Empirical Finance*

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

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

Lamoureux, C., and W. Lastrapes (1990) ‘Persistence in Variance, Strucutral Change, and the GARCH Model.’ *Journal of Business and Economic Statistics* 8, 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

Lopez, J. A. (2001) ‘Evaluating the Predictive Accuracy of Volatility Models.’ *Journal of Forecasting* 20(1), 87–109

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

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 Romano J. P. (1994) ‘The Stationary Bootstrap.’ *Journal of The American Statistical Association* 89, 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, 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, 23–45

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

Table 1: Descriptive Statistics of r_t

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 October 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.

Table 2: Maximum Likelihood Estimates of Standard GARCH Models with different conditional distributions.

	GARCH-N	GARCH-t	GARCH-GED	EGARCH-N	EGARCH-t	EGARCH-GED	GJR-N	GJR-t	GJR-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. Asymptotic standard errors are in parentheses.

Table 3: Maximum Likelihood Estimates of MRS-GARCH Models with different conditional distributions.

	MRS-GARCH-N	MRS-GARCH-t2	MRS-GARCH-t	MRS-GARCH-GED
$\delta^{(1)}$	0.0592	0.0571	0.0487	0.0715
	(0.0140)	(0.0140)	(0.0130)	(0.0210)
$\delta^{(2)}$	-1.6623	0.0558	0.0792	0.0252
	(0.2090)	(0.0310)	(0.0290)	(0.0120)
$\alpha_0^{(1)}$	0.0062	0.0026	0.0359	0.0923
	(0.0040)	(0.0010)	(0.0090)	(0.0180)
$\alpha_0^{(2)}$	0.5961	0.1132	0.1130	0.0341
	(0.1420)	(0.0310)	(0.0210)	(0.0100)
$\alpha_1^{(1)}$	0.0230	0.0137	0.0413	0.0560
	(0.0070)	(0.0050)	(0.0120)	(0.0140)
$\alpha_1^{(2)}$	0.0225	0.0754	0.0565	0.0388
	(0.1170)	(0.0190)	(0.0160)	(0.0150)
$\beta_1^{(1)}$	0.9093	0.9805	0.8473	0.8952
	(0.0090)	(0.0060)	(0.0330)	(0.0200)
$\beta_1^{(2)}$	0.9633	0.8569	0.8924	0.8533
	(0.1810)	(0.0290)	(0.0210)	(0.0370)
p	0.9811	0.9987	0.9998	0.9999
	(0.0040)	(0.0010)	(0.0001)	(0.0001)
q	0.1533	0.9988	0.9999	0.9998
	(0.0850)	(0.0010)	(0.0001)	(0.0002)
$\nu^{(1)}$	-	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	0.02	0.52	0.61	0.33
π_2	0.98	0.48	0.39	0.67
d_1	53.01	765.70	4901.96	8771.93
d_2	1.18	842.46	7518.80	4291.85

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 10/28/2001. The superscripts indicate the regime. π_j is the unconditional probability of being in regime j, while d_j is the half-life or expected duration of the j-th state. Asymptotic standard errors are in parentheses.

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.

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

1-step ahead volatility forecasts

Model	MSE_1	Rank	$MSE2$	Rank	$QLIKE$	Rank	$R2LOG$	Rank	$MAD2$	Rank	$MAD1$	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	$MSE2$	Rank	$QLIKE$	Rank	$R2LOG$	Rank	$MAD2$	Rank	$MAD1$	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: Out-of-sample evaluation of one- and five-step ahead volatility forecasts. The volatility proxy is given by the realized volatility at 1-minute.

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

10-step ahead volatility forecasts

Model	MSE_1	Rank	$MSE2$	Rank	$QLIKE$	Rank	$R2LOG$	Rank	$MAD2$	Rank	$MAD1$	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	$MSE2$	Rank	$QLIKE$	Rank	$R2LOG$	Rank	$MAD2$	Rank	$MAD1$	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: Out-of-sample evaluation of one- and five-step ahead volatility forecasts. The volatility proxy is given by the realized volatility at 1-minute.

Table 7: **Diebold-Mariano Test.** (Benchmark: MRS-GARCH-N, 1-step-ahead)

Model	MSE_1	$MSE2$	$QLIKE$	$R2LOG$	$MAD2$	$MAD1$	$HMSE$
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. †and ††represent the DM statistics for which one can reject the null at 5% and 1% respectively, but the sign of the statistics is positive, indicating that the benchmark implies a bigger loss.

Table 8: **Diebold-Mariano Test.** (Benchmark: EGARCH-N, 10-step-ahead)

Model	MSE_1	$MSE2$	$QLIKE$	$R2LOG$	$MAD2$	$MAD1$	$HMSE$
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. †and ††represent the DM statistics for which one can reject the null at 5% and 1% respectively, but the sign of the statistics is positive, indicating that the benchmark implies a bigger loss.

Table 9: **Reality Check and Superior Predictive Ability Tests** (all models, one-day horizon).

		Loss Functions						
Benchmark		MSE_1	$MSE2$	$QLIKE$	$R2LOG$	$MAD1$	$MAD2$	$HMSE$
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 Consistent (SPA_c^0) and Lower bound (SPA_L^0) Hansen's (2001) SPA test of 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 are better than the benchmark. The number of bootstrap replications to calculate the p-values is 3000 and the block length is 0.33.

Table 10: **Reality Check and Superior Predictive Ability Tests** (MRS-GARCH models only, two-week horizon).

		Loss Functions						
Benchmark		MSE_1	$MSE2$	$QLIKE$	$R2LOG$	$MAD1$	$MAD2$	$HMSE$
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 Consistent (SPA_c^0) and Lower bound (SPA_L^0) Hansen's (2001) SPA test of forecasts at 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 are better than the benchmark. The number of bootstrap replications to calculate the p-values is 3000 and the block length is 0.3.

Table 11: **Risk management Out-of-sample Evaluation: 95% and 99% VaR**

95% VaR

Steps	1						5						10						22					
Model	TUFF	PF(%)	Rank	**LRPF**	**LRind**	**LRcc**	TUFF	PF(%)	Rank	**LRPF**	**LRind**	**LRcc**	TUFF	PF(%)	Rank	**LRPF**	**LRind**	**LRcc**	TUFF	PF(%)	Rank	**LRPF**	**LRind**	**LRcc**
GARCH-N	20	6.458	11=	2.102	0.010	2.112	49	5.871	11=	0.775	34.293*	35.067*	77	5.088	10	0.008	42.734*	42.743*	69	9.002	9=	14.070*	135.694*	149.764*
GARCH-t	20	3.718	3=	1.932	0.116	2.049	86	2.544	3=	7.854*	21.812*	29.665*	186	1.566	2	17.148*	24.637*	41.784*	70	4.501	4	0.277	76.168*	76.445*
GARCH-GED	20	6.458	11=	2.102	0.010	2.112	49	5.871	11=	0.775	34.293*	35.067*	77	5.284	11	0.085	47.098*	47.183*	69	9.002	9=	14.070*	135.694*	149.764*
EGARCH-N	20	4.892	8	0.013	2.579	2.591	49	4.892	8=	0.013	26.224*	26.236*	77	5.479	12	0.240	44.659*	44.899*	65	13.894	12	58.625*	156.722*	215.347*
EGARCH-t	20	4.697	7	0.101	2.443	2.544	49	4.892	8=	0.013	32.066*	32.079*	169	4.892	8=	0.013	60.188*	60.201*	129	8.806	7=	12.832*	131.663*	144.496*
EGARCH-GED	20	3.718	3=	1.932	1.471	3.403	49	4.305	6	0.544	25.474*	26.018*	77	4.892	8=	0.013	38.380*	38.392*	65	12.916	11	47.839*	141.262*	189.101*
GJR-N	20	5.675	10	0.471	3.499	3.970	49	4.892	8=	0.013	20.859*	20.871*	77	3.718	5=	1.932	24.513*	26.445*	69	8.806	7=	12.832*	114.588*	127.420*
GJR-t	20	3.718	3=	1.932	1.471	3.403	49	2.544	3=	7.854*	4.319*	12.173*	187	1.761	3=	14.876*	22.123*	36.999*	70	4.11	3	0.906	48.444*	49.350*
GJR-GED	20	5.479	9	0.240	3.255	3.495	49	4.501	7	0.277	18.460*	18.737*	77	3.718	5=	1.932	24.513*	26.445*	69	8.611	6	11.643*	118.947*	130.590*
MRS-GARCH-N	20	7.045	13	4.014*	0.092	4.106	41	7.241	13	4.773*	32.357*	37.130*	70	8.611	13	11.643*	67.578*	79.221*	65	17.417	13	103.924*	261.995*	365.918*
MRS-GARCH-t2	86	1.761	1	14.876*	2.164	17.041*	200	0.978	1	25.646*	23.353*	48.999*	200	0.587	1	33.279*	18.523*	51.802*	175	1.761	1	14.876*	42.362*	57.239*
MRS-GARCH-t	20	2.544	2	7.854*	0.945	8.798*	187	1.37	2	19.674*	17.739*	37.413*	187	1.761	3=	14.876*	14.026*	28.903*	175	3.327	2	3.397	62.162*	65.559*
MRS-GARCH-GED	20	4.305	6	0.544	0.989	1.533	49	3.914	5	1.367	35.952*	37.319*	77	4.11	7	0.906	33.734*	34.640*	69	8.415	5	10.503*	114.918*	125.422*

99% VaR

Steps	1						5						10						22					
Model	TUFF	PF(%)	Rank	**LRPF**	**LRind**	**LRcc**	TUFF	PF(%)	Rank	**LRPF**	**LRind**	**LRcc**	TUFF	PF(%)	Rank	**LRPF**	**LRind**	**LRcc**	TUFF	PF(%)	Rank	**LRPF**	**LRind**	**LRcc**
GARCH-N	20	2.544	13	8.621*	0.945	9.566*	206	0.587	9	1.033	6.844*	7.878*	187	0.783	7	0.263	15.083*	15.346*	175	1.761	8=	2.438	42.362*	44.801*
GARCH-t	86	0.391	3	2.487	0.016	2.503	207	0.196	2=	4.991*	0.004	4.995	200	0.587	1=	1.033	18.523*	19.556*	188	0.587	4	1.033	18.523*	19.556*
GARCH-GED	86	1.174	8	0.148	0.143	0.291	206	0.391	5=	2.487	8.926*	11.412*	200	0.587	1=	1.033	18.523*	19.556*	187	0.978	6	0.002	23.353*	23.355*
EGARCH-N	20	1.566	10	1.408	0.255	1.663	187	0.783	10	0.263	5.505*	5.768	187	1.566	12	1.408	24.637*	26.045*	70	5.871	12	57.665*	92.787*	150.453*
EGARCH-t	86	0.783	6=	0.263	0.063	0.326	200	1.174	12	0.148	20.197*	20.346*	187	1.174	10	0.148	10.944*	11.093*	187	1.761	8=	2.438	42.362*	44.801*
EGARCH-GED	86	0.587	4=	1.033	0.036	1.069	206	0.391	5=	2.487	8.926*	11.412*	200	0.978	8=	0.002	23.353*	23.355*	129	2.153	11	5.156*	56.608*	61.764*
GJR-N	20	1.761	11	2.438	0.323	2.762	187	0.978	11	0.002	4.522*	4.525	187	1.37	11	0.633	17.739*	18.372*	175	1.761	8=	2.438	42.362*	44.801*
GJR-t	86	0.587	4=	1.033	0.036	1.069	511	0	1	10.271*	0.000	10.271*	200	0.587	1=	1.033	18.523*	19.556*	189	0.196	2	4.991*	0.004	4.995
GJR-GED	20	1.37	9	0.633	0.195	0.828	206	0.391	5=	2.487	8.926*	11.412*	200	0.978	8=	0.002	23.353*	23.355*	187	1.174	7	0.148	31.493*	31.641*
MRS-GARCH-N	20	2.348	12	6.803*	0.578	7.382*	187	1.566	13	1.408	15.727*	17.135*	186	2.153	13	5.156*	26.028*	31.185*	69	7.828	13	97.298*	119.832*	217.130*
MRS-GARCH-t2	511	0	1=	10.271*	0.000	10.271*	207	0.196	2=	4.991*	0.004	4.995	200	0.587	1=	1.033	18.523*	19.556*	511	0	1	10.271*	0.000	10.271*
MRS-GARCH-t	511	0	1=	10.271*	0.000	10.271*	207	0.196	2=	4.991*	0.004	4.995	200	0.587	1=	1.033	18.523*	19.556*	189	0.391	3	2.487	8.926*	11.412*
MRS-GARCH-GED	86	0.783	6=	0.263	0.063	0.326	206	0.391	5=	2.487	8.926*	11.412*	200	0.587	1=	1.033	18.523*	19.556*	187	0.783	5	0.263	27.804*	28.067*

Note: This table presents the time until first failure (TUFF), the percentage proportion of failures (PF(%)), the LR test for unconditional coverage (LRPF), the LR test for independence (LRind), and the LR test for conditional coverage (LRcc) for both 95% and 99% VaR failure processes at one, five, ten and twenty-two steps ahead. * indicates significance at 5%.

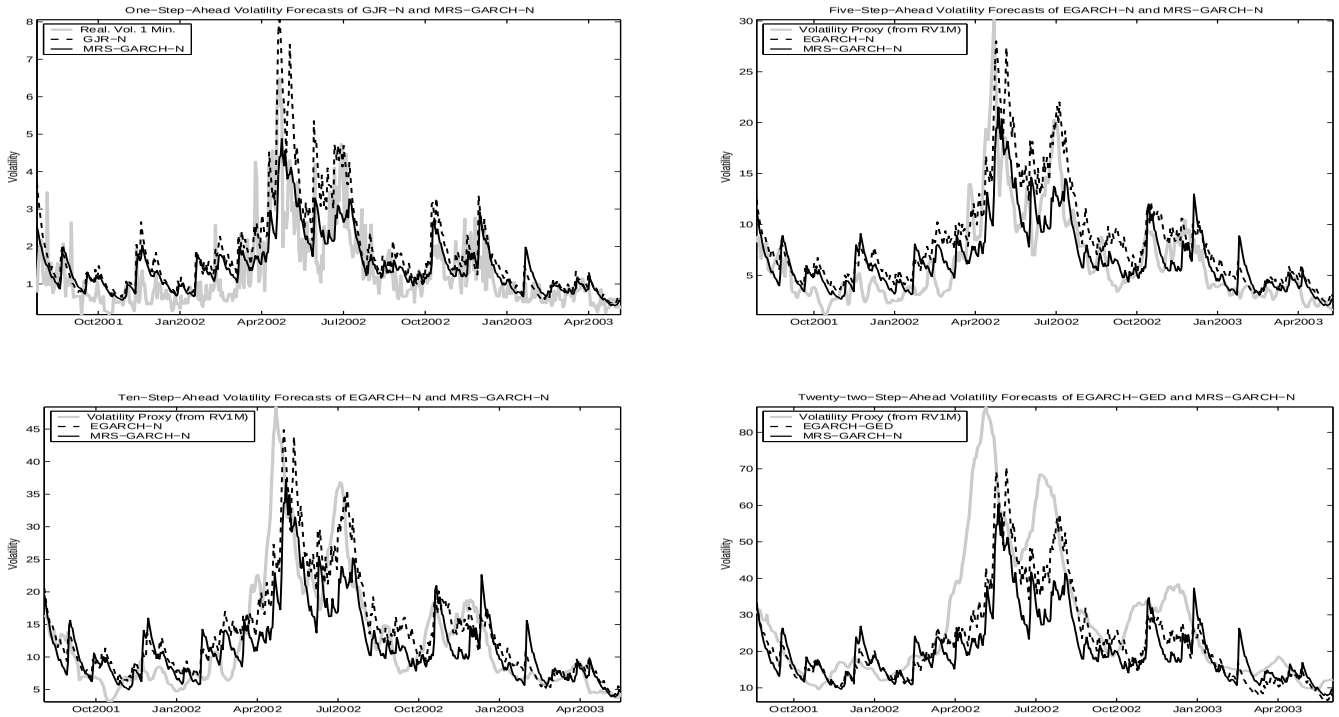

Figure 1: Comparison of one, five, ten and twenty-step-ahead volatility forecasts from the Markov Regime-Switching GARCH Models and standard GARCH.

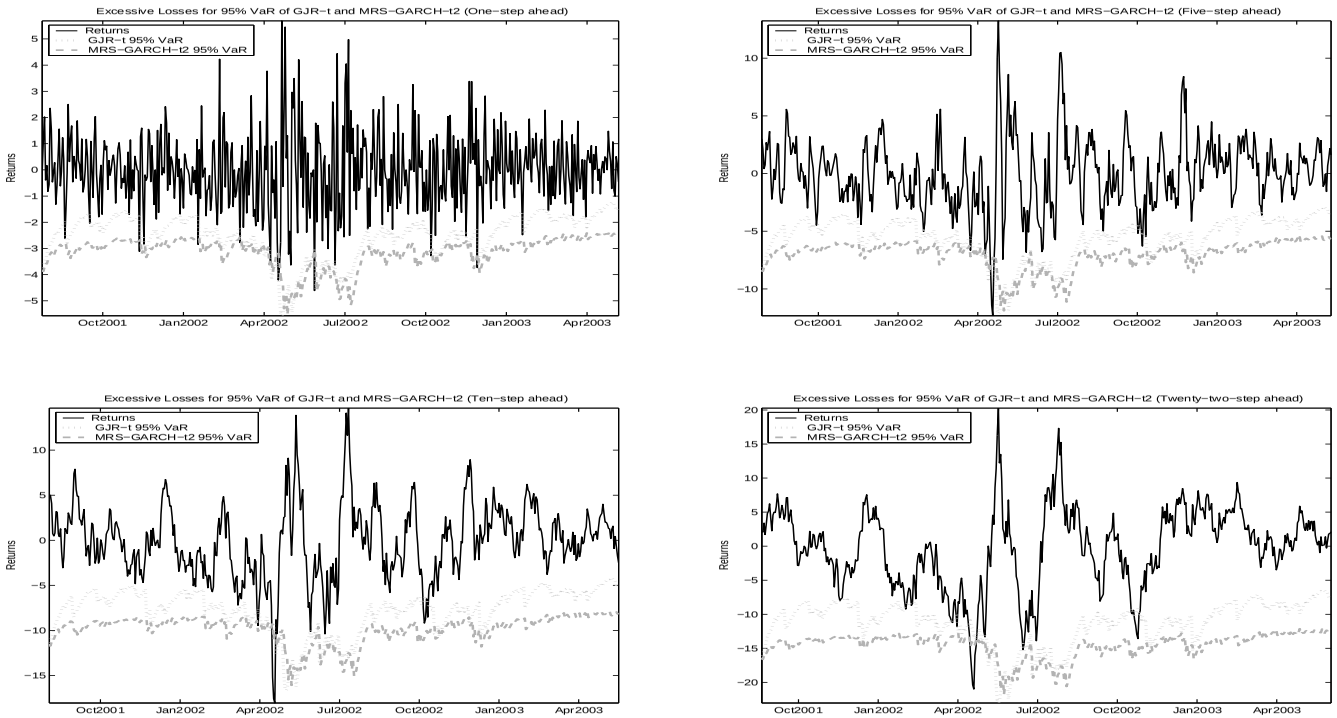

Figure 2: 95% VaR estimates for S&P100 series.

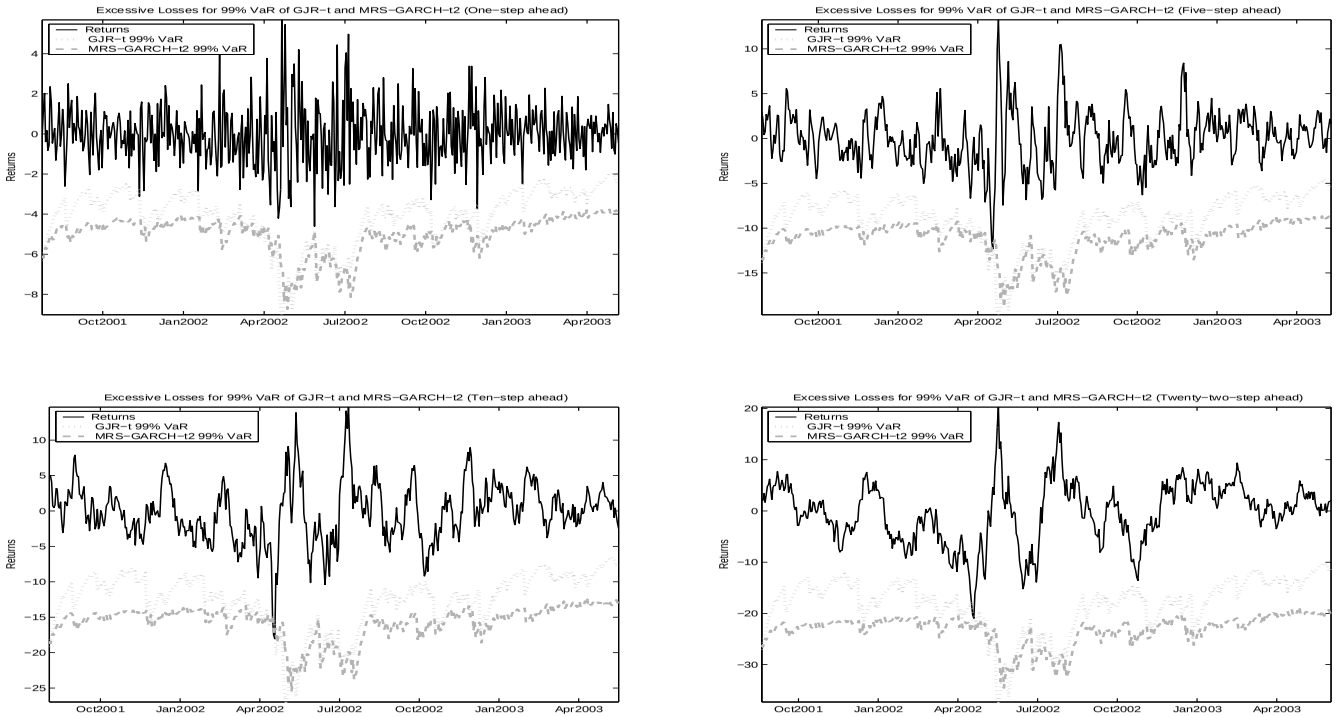

Figure 3: 99% VaR estimates for S&P100 series.