JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **20**: 873–889 (2005)
Published online 30 March 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jae.800

A FORECAST COMPARISON OF VOLATILITY MODELS: DOES ANYTHING BEAT A GARCH(1,1)?

PETER R. HANSEN[a]* AND ASGER LUNDE[b]

[a] *Department of Economics, Brown University, Providence, USA*
[b] *Department of Information Science, Aarhus School of Business, Denmark*

SUMMARY

We compare 330 ARCH-type models in terms of their ability to describe the conditional variance. The models are compared out-of-sample using DM–$ exchange rate data and IBM return data, where the latter is based on a new data set of realized variance. We find no evidence that a GARCH(1,1) is outperformed by more sophisticated models in our analysis of exchange rates, whereas the GARCH(1,1) is clearly inferior to models that can accommodate a leverage effect in our analysis of IBM returns. The models are compared with the test for superior predictive ability (SPA) and the reality check for data snooping (RC). Our empirical results show that the RC lacks power to an extent that makes it unable to distinguish 'good' and 'bad' models in our analysis. Copyright © 2005 John Wiley & Sons, Ltd.

1. INTRODUCTION

The conditional variance of financial time series is important for pricing derivatives, calculating measures of risk, and hedging. This has sparked an enormous interest in modelling the conditional variance and a large number of volatility models have been developed since the seminal paper of Engle (1982); see Poon and Granger (2003) for an extensive review and references.

The aim of this paper is to examine whether sophisticated volatility models provide a better description of financial time series than more parsimonious models. We address this question by comparing 330 GARCH-type models in terms of their ability to forecast the one-day-ahead conditional variance. The models are evaluated out-of-sample using six different loss functions, where the realized variance is substituted for the latent conditional variance. We use the test for superior predictive ability (SPA) of Hansen (2001) and the reality check for data snooping (RC) by White (2000) to benchmark the 330 volatility models to the GARCH(1,1) of Bollerslev (1986). These tests have the advantage that they properly account for the full set of models, without the use of probability inequalities, such as the Bonferroni bound, that typically lead to conservative tests.

We compare the models using daily DM–$ exchange rate data and daily IBM returns. There are three main findings of our empirical analysis. First, in the analysis of the exchange rate data we find no evidence that the GARCH(1,1) is inferior to other models, whereas the GARCH(1,1) is clearly outperformed in the analysis of IBM returns. Second, our model space includes models with many distinct characteristics that are interesting to compare,[1] and some interesting details

* Correspondence to: Peter R. Hansen, Brown University, Department of Economics, Box B, Providence, RI 02 912, USA. E-mail: peter_hansen@brown.edu

[1] However, we have by no means included all the volatility models that have been proposed in the literature. For a comparison of a smaller set of models that also includes stochastic volatility models and fractionally integrated GARCH models, see Hansen *et al*. (2003).

Received 8 November 2002
Revised 12 February 2004

emerge from the out-of-sample analysis. The models that perform well in the IBM return data are primarily those that can accommodate a leverage effect, and the best overall performance is achieved by the A-PARCH(2,2) model of Ding *et al*. (1993). Other aspects of the volatility models are more ambiguous. While the *t*-distributed specification of standardized returns generally leads to a better average performance than the Gaussian in the analysis of exchange rates, the opposite is the case in our analysis of IBM returns. The different mean specifications, zero-mean, constant mean and GARCH-in-mean, result in almost identical performances. Third, our empirical analysis shows that the RC has less power than the SPA test. This makes an important difference in our application, because the RC cannot detect that the GARCH(1,1) is significantly outperformed by other models in the analysis of IBM returns. In fact, the RC even suggests that an ARCH(1) may be the best model in many cases, which does not conform with the existing empirical evidence. The SPA test always finds the ARCH(1) model to be inferior, which shows that the SPA test has power in these applications and is therefore more likely to detect superior models when such exist.

Ideally, we would evaluate the models' ability to forecast all aspects of the conditional distribution. However, it is not possible to extract precise information about the conditional distribution without making restrictive assumptions. Instead we focus on the central component of the models—the conditional variance—that can be estimated by the realized variance. Initially, it was common to substitute the squared return for the unobserved conditional variance in out-of-sample evaluations of volatility models. This typically resulted in a poor performance, which instigated a discussion of the practical relevance of volatility models. However, Andersen and Bollerslev (1998) showed that the 'poor' performance could be explained by the fact that the squared return is a noisy proxy for the conditional variance. By substituting the realized variance (instead of the squared return), Andersen and Bollerslev (1998) showed that volatility models perform quite well. Hansen and Lunde (2003) provide another important argument for using the realized variance rather than the squared return. They show that substituting the squared returns for the conditional variance can severely distort the comparison, in the sense that the empirical ranking of models may be inconsistent for the true (population) ranking. So an evaluation that is based on squared returns may select an inferior model as the 'best' with a probability that converges to one as the sample size increases. For this reason, our evaluation is based on the realized variance.

Comparing multiple models is a non-standard inference problem, and spurious results are likely to appear unless inference controls for the multiple comparisons. An inferior model can be 'lucky' and perform better than all other models, and the more models that are being compared the higher is the probability that the best model (in population) has a much smaller sample performance than some inferior model. It is therefore important to control for the full set of models and their interdependence when evaluating the significance of an excess performance. In our analysis we employ the SPA test and the RC, which are based on the work of Diebold and Mariano (1995) and West (1996). These tests can evaluate whether a particular model (benchmark) is significantly outperformed by other models, while taking into account the large number of models that are being compared. In other words, these tests are designed to evaluate whether an observed excess performance is significant or could have occurred by chance.

This paper is organized as follows. Section 2 describes the 330 volatility models under consideration and the loss functions are defined in Section 3. In Section 4, we describe our measures of realized variance and Section 5 contains some details of the SPA test and its bootstrap implementation. We present our empirical results in Section 6 and Section 7 contains some concluding remarks.

2. THE GARCH UNIVERSE

Given a price process, p_t, we define the compounded daily return by $r_t = \log(p_t) - \log(p_{t-1})$, $t = -R+1, \ldots, n$. Later we split the sample into an estimation period (the first R observations) and an evaluation period (the last n observations).

The conditional density of r_t is denoted by $f(r|\mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1}$ is the σ-algebra induced by variables that are observed at time $t-1$. We define the conditional mean by $\mu_t \equiv E(r_t|\mathcal{F}_{t-1})$ (the *location* parameter) and the conditional variance by $\sigma_t^2 \equiv \text{var}(r_t|\mathcal{F}_{t-1})$ (the *scale* parameter), assuming that both are well defined. Subsequently we can define the standardized return, $e_t \equiv (r_t - \mu_t)/\sigma_t$, and denote its conditional density by $g(e|\mathcal{F}_{t-1})$. Following Hansen (1994) we consider a parametric specification, $f(r|\psi(\mathcal{F}_{t-1};\theta))$, where $\theta \in \Theta \subset \mathbb{R}^q$ is a vector of parameters. It now follows that the *time-varying* vector of parameters, $\psi_t \equiv \psi(\mathcal{F}_{t-1};\theta)$, can be divided into $\psi_t = (\mu_t, \sigma_t^2, \eta_t)$, where η_t is a vector of *shape* parameters for the conditional density of e_t. Thus, we have a family of density functions for r_t, which is a location-scale family with (possibly time-varying) shape parameters, and we shall model μ_t, σ_t^2 and η_t individually. Most GARCH-type models can be formulated in this framework and η_t typically does not depend on t.

The notation for our modelling of the conditional mean and variance is $m_t = \mu(\mathcal{F}_{t-1};\theta)$ and $h_t^2 = \sigma^2(\mathcal{F}_{t-1};\theta)$, respectively, and we employ two specifications for $g(e|\eta_t)$ in our empirical analysis. One is a Gaussian specification that is free of parameters $g(e|\eta_t) = g(e)$, and the other is a t-specification that has degrees of freedom, υ, as the only parameter, $g(e|\eta_t) = g(e|\upsilon)$.[2] Our specifications for the conditional mean are: $m_t = \mu_0 + \mu_1\sigma_{t-1}^2$ (GARCH-in-mean), $m_t = \mu_0$ and $m_t = 0$.

The conditional variance is the main object of interest and our analysis includes a large number of parametric specifications for σ_t that are listed in Table I. The use of acronyms has not been fully consistent in the existing literature, for example, A-GARCH has been used to represent four different specifications. So to avoid any confusion we use 'A-GARCH' to refer to a model by Engle and Ng (1993) and use different acronyms for all other models, e.g., we use H-GARCH to refer to the model by Hentshel (1995). Several specifications nest other specifications, as is evident from Table I. In particular, the flexible specifications of the H-GARCH and the Aug-GARCH, see Duan (1997), nest many of the simpler specifications. An empirical comparison of several of the models that are nested in the Aug-GARCH model can be found in Loudon *et al*. (2000).

The evolution of volatility models has been motivated by empirical findings and economic interpretations. Ding *et al*. (1993) used Monte Carlo simulations to demonstrate that both the GARCH specification (model for σ_t^2) and the TS-GARCH specification[3] (model for σ_t) are capable of producing the autocorrelation pattern that is seen in financial data. So in this respect there is no argument for modelling σ_t rather than σ_t^2 or vice versa. More generally, we can consider a modelling of σ_t^δ where δ is a parameter to be estimated, and this motivated the *Box–Cox transformations* that involve σ_t and ε_t. The empirically observed *leverage effect* motivated the development of models with an asymmetric response in volatility to positive and negative shocks. The leverage effect was first noted by Black (1976) and is best illustrated by the *news impact curve*, which was introduced by Pagan and Schwert (1990) and named by Engle and Ng (1993). This curve is a plot of σ_t^2 against ε_{t-1} that illustrates how the volatility reacts to good and bad news.

In our analysis, we have included the four combinations of p, $q = 1, 2$ for the lag length parameters, with the following exceptions: the ARCH is only estimated for $q = 1$; H-GARCH

[2] We do not restrict υ to be an integer.
[3] See Taylor (1986) and Schwert (1990).

Table I. Specifications for the conditional variance

ARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2$
GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
IGARCH	$\sigma_t^2 = \omega + \varepsilon_{t-1}^2 + \sum_{i=2}^{q} \alpha_i (\varepsilon_{t-i}^2 - \varepsilon_{t-1}^2) + \sum_{j=1}^{p} \beta_j (\sigma_{t-j}^2 - \varepsilon_{t-1}^2)$
Taylor/Schwert:	$\sigma_t = \omega + \sum_{i=1}^{q} \alpha_i \lvert\varepsilon_{t-i}\rvert + \sum_{j=1}^{p} \beta_j \sigma_{t-j}$
A-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} [\alpha_i \varepsilon_{t-i}^2 + \gamma_i \varepsilon_{t-i}] + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
NA-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i (\varepsilon_{t-i} + \gamma_i \sigma_{t-i})^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
V-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i (e_{t-i} + \gamma_i)^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
Thr.-GARCH:	$\sigma_t = \omega + \sum_{i=1}^{q} \alpha_i [(1-\gamma_i)\varepsilon_{t-i}^{+} - (1+\gamma_i)\varepsilon_{t-i}^{-}] + \sum_{j=1}^{p} \beta_j \sigma_{t-j}$
GJR-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} [\alpha_i + \gamma_i I_{(\varepsilon_{t-i}>0)}] \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
log-GARCH:	$\log(\sigma_t) = \omega + \sum_{i=1}^{q} \alpha_i \lvert e_{t-i}\rvert + \sum_{j=1}^{p} \beta_j \log(\sigma_{t-j})$
EGARCH:	$\log(\sigma_t^2) = \omega + \sum_{t=1}^{q} [\alpha_i e_{t-i} + \gamma_i (\lvert e_{t-i}\rvert - E\lvert e_{t-i}\rvert)] + \sum_{j=1}^{p} \beta_j \log(\sigma_{t-j}^2)$
NGARCH:[a]	$\sigma_t^{\delta} = \omega + \sum_{i=1}^{q} \alpha_i \lvert\varepsilon_{t-i}\rvert^{\delta} + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^{\delta}$
A-PARCH:	$\sigma^{\delta} = \omega + \sum_{i=1}^{q} \alpha_i [\lvert\varepsilon_{t-i}\rvert - \gamma_i \varepsilon_{t-i}]^{\delta} + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^{\delta}$
GQ-ARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i} + \sum_{i=1}^{p} \alpha_{ii} \varepsilon_{t-i}^2 + \sum_{i<j}^{p} \alpha_{ij} \varepsilon_{t-i} \varepsilon_{t-j} + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
H-GARCH:	$\sigma_t^{\delta} = \omega + \sum_{i=1}^{q} \alpha_i \delta \sigma_{t-i}^{\delta} [\lvert e_t - \kappa\rvert - \tau(e_t - \kappa)]^{\nu} + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^{\delta}$
Aug-GARCH:[b]	$\sigma_t^2 = \begin{cases} \lvert\delta\phi_t - \delta + 1\rvert^{1/\delta} & \text{if } \delta \neq 0 \\ \exp(\phi_t - 1) & \text{if } \delta = 0 \end{cases}$
	$\phi_t = \omega + \sum_{i=1}^{q} [\alpha_{1i} \lvert\varepsilon_{t-i} - \kappa\rvert^{\nu} + \alpha_{2i} \max(0, \kappa - \varepsilon_{t-i})^{\nu}] \phi_{t-j}$
	$+ \sum_{i=1}^{q} [\alpha_{3i} f(\lvert\varepsilon_{t-i} - \kappa\rvert, \nu) + \alpha_{4i} f(\max(0, \kappa - \varepsilon_{t-i}), \nu)] \phi_{t-j}$
	$+ \sum_{j=1}^{p} \beta_j \phi_{t-j}^2$

[a] This is the A-PARCH model without the leverage effect.
[b] Here $f(x, \nu) = (x^{\nu} - 1)/\nu$.

and Aug-GARCH are only estimated for $(p, q) = (1, 1)$, because these are quite burdensome to estimate. It is well known that an ARCH(1) model is unable to fully capture the persistence in volatility, and this model is only included as a point of reference, and to verify that the tests, SPA and RC, have power. This is an important aspect of the analysis, because a test that is unable to reject that the ARCH(1) is the best model cannot be very informative about which is a better model. Restricting the models to have two lags (or less) should not affect the main conclusions of our empirical analysis, because it is unlikely that a model with more lags would outperform a simple benchmark in the out-of-sample comparison, unless the same model with two lags can outperform the benchmark. This aspect is also evident from our analysis, where a model with $p = q = 2$ rarely performs better (out-of-sample) than the same model with fewer lags, even though most parameters are found to be significant (in-sample).

3. FORECAST EVALUATION

A popular way to evaluate volatility models out-of-sample is in terms of the R^2 from a Mincer–Zarnowitz (MZ) regression, $r_t^2 = a + bh_t^2 + u_t$, where squared returns are regressed on the model forecast of σ_t^2 and a constant. Or the logarithmic version, $\log(r_t^2) = a + b\log(h_t^2) + u_t$, that is less sensitive to outliers, as was noted by Pagan and Schwert (1990) and Engle and Patton (2001).[4] However, the R^2 of a MZ regression is not an ideal criterion for comparing volatility models, because it does not penalize a biased forecast. For example, a poor biased forecast may achieve a higher R^2 than a good unbiased forecast, because the bias can be eliminated artificially through estimates of (a, b) that differ from (0, 1).

It is not obvious which loss function is more appropriate for the evaluation of volatility models, as discussed by Bollerslev *et al*. (1994), Diebold and Lopez (1996) and Lopez (2001). So rather than making a single choice we use the following six loss functions in our empirical analysis:

$$\mathrm{MSE}_1 \equiv n^{-1}\sum_{t=1}^{n}(\sigma_t - h_t)^2 \qquad \mathrm{MSE}_2 \equiv n^{-1}\sum_{t=1}^{n}(\sigma_t^2 - h_t^2)^2$$

$$\mathrm{QLIKE} \equiv n^{-1}\sum_{t=1}^{n}(\log(h_t^2) + \sigma_t^2 h_t^{-2}) \qquad \mathrm{R^2LOG} \equiv n^{-1}\sum_{t=1}^{n}[\log(\sigma_t^2 h_t^{-2})]^2$$

$$\mathrm{MAE}_1 \equiv n^{-1}\sum_{t=1}^{n}|\sigma_t - h_t| \qquad \mathrm{MAE}_2 \equiv n^{-1}\sum_{t=1}^{n}|\sigma_t^2 - h_t^2|$$

The criteria MSE_2 and $\mathrm{R^2LOG}$ are similar to the R^2 of the MZ regressions,[5] and QLIKE corresponds to the loss implied by a Gaussian likelihood. The mean absolute error criteria, MAE_2 and MAE_1, are interesting because they are more robust to outliers than, say, MSE_2. Additional discussions of the MSE_2, QLIKE and $\mathrm{R^2LOG}$ criteria can be found in Bollerslev *et al*. (1994).

[4] Engle and Patton (2001) also point out that heteroskedastic returns imply (even more) heteroskedasticity in the squared returns, r_t^2. So parameters are estimated inefficiently and the usual standard errors are misleading.

[5] Provided that $a = 0$ and $b = 1$, which essentially requires the forecasts to be unbiased.

4. REALIZED VARIANCE

In our empirical analysis we substitute the realized variance for the latent σ_t^2. The realized variance for a particular day is calculated from intraday returns, $r_{i,m}$, where $r_{t,i,m} \equiv p_{t-(i-1)/m} - p_{t-i/m}$ for $i = 1, \ldots, m$. Thus $r_{t,i,m}$ is the return over a time interval with length $1/m$ on day t, and we note that $r_t = \sum_{i=1}^m r_{t,i,m}$. It will often be reasonable to assume that $E(r_{t,i,m}|\mathcal{F}_{t-1}) \simeq 0$ and that intraday returns are conditionally uncorrelated, $\operatorname{cov}(r_{t,i,m}, r_{t,j,m}|\mathcal{F}_{t-1}) = 0$ for $i \neq j$, such that $\sigma_t^2 \equiv \operatorname{var}\left(\sum_{i=1}^m r_{t,i,m}|\mathcal{F}_{t-1}\right) = \sum_{i=1}^m \operatorname{var}(r_{t,i,m}|\mathcal{F}_{t-1}) \simeq \sum_{i=0}^m E(r_{t,i,m}^2|\mathcal{F}_{t-1}) = E[RV_t^{(m)}|\mathcal{F}_{t-1}]$, where we have defined the realized variance (at frequency m) $RV_t^{(m)} \equiv \sum_{i=1}^m r_{t,i,m}^2$. Thus $RV_t^{(m)}$ is approximately unbiased for σ_t^2 (given our assumptions above), and it can often be shown that $E[RV_t^{(m)} - \sigma_t^2]^2$ is decreasing in m, such that $RV_t^{(m)}$ is an increasingly more precise estimator of σ_t^2 as m increases.[6] Further, the $RV_t^{(m)}$ is (by definition) consistent for the quadratic variation of p_t, which is identical to the conditional variance, σ_t^2, for certain data generating processes (DGPs) such as the ARCH-type models considered in this paper.[7]

Several assets are not traded 24 hours a day, because the market is closed overnight and over weekends. In these situations we only observe $f \leq m$ (of the m possible) intraday returns. Assume for simplicity that we observe, $r_{t,1,m}, \ldots, r_{t,f,m}$ and define $RV_t^{(f/m)} \equiv \sum_{i=1}^f r_{t,i,m}^2$. Since $RV_t^{(f/m)}$ only captures the volatility during the part of the day that the market is open, we need to extend $RV_t^{(f/m)}$ to a measure of volatility for the full day. One resolution is to add the squared close-to-open return to $RV_t^{(f/m)}$, but this leads to a noisy measure because 'overnight' returns are relatively noisy. A better solution is to scale $RV_t^{(f/m)}$, and use the estimator

$$\hat{\sigma}_t^2 \equiv \hat{c} \cdot RV_t^{(f/m)} \quad \text{where} \quad \hat{c} \equiv \left(\frac{n^{-1} \sum_{t=1}^n (r_t - \hat{\mu}_t)^2}{n^{-1} \sum_{t=1}^n RV_t^{(f/m)}} \right) \tag{1}$$

This yields an estimator that is approximately unbiased for σ_t^2 under fairly reasonable assumptions. See Martens (2002), Hol and Koopman (2002) and Fleming *et al*. (2003), who applied similar scaling estimators to obtain a measure of volatility for the whole day.

5. TEST FOR SUPERIOR PREDICTIVE ABILITY

We divide the observations into an estimation period and an evaluation period:

$$t = \underbrace{-R+1, \ldots, 0}_{\text{estimation period}}, \quad \underbrace{1, 2, \ldots, n}_{\text{evaluation period}}$$

[6] In practice, m must be chosen moderately large, to avoid that intraday returns become correlated due to market microstructure effects. In the technical appendix (Hansen and Lunde, 2001), we list the R^2 values from two MZ regressions, $r_t^2 = a + bh_t^2 + u_t$ and $RV_t^{(288)} = a + bh_t^2 + u_t$, where the realized variance, $RV_t^{(288)}$, is defined in the next section. The R^2 of the former typically lies between 2 and 4%, whereas the R^2 of the latter lies between 35 and 45%. This strongly suggests that $RV_t^{(288)}$ is a far more precise estimate of σ_t^2 than is r_t^2.

[7] For other DGPs the $RV_t^{(m)}$ is consistent for the integrated variance, see Meddahi (2002) and Barndorff-Nielsen and Shephard (2001), which need not equal σ_t^2. However, this does not change our main argument for using the realized variance, which is that $RV_t^{(m)}$ is a more precise estimator of σ_t^2 than is r_t^2.

The parameters of the volatility models are estimated using the first R interday observations, and these estimates are used to make one-step-ahead forecasts for the remaining n periods. During the evaluation period we calculate the realized variance from intraday returns and obtain $\hat{\sigma}_t^2$ using (1). Thus model k yields a sequence of forecasts, $h_{k,1}^2, \ldots, h_{k,n}^2$, that are compared to $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_n^2$, using a loss function L. Let the first model, $k = 0$, be the benchmark model that is compared to models $k = 1, \ldots, l$. Each model leads to a sequence of losses, $L_{k,t} \equiv L(\hat{\sigma}_t^2, h_{k,t}^2)$, $t = 1, \ldots, n$, and we define the relative performance variables

$$X_{k,t} \equiv L_{0,t} - L_{k,t}, \quad k = 1, \ldots, l, t = 1, \ldots, n$$

Our null hypothesis is that the benchmark model is as good as any other model in terms of expected loss. This can be formulated as the hypothesis $H_0 : \lambda_k \equiv E(X_{k,t}) \leq 0$ for all $k = 1, \ldots, l$, because $\lambda_k > 0$ corresponds to the case where model k is better than the benchmark. In order to apply the stationary bootstrap of Politis and Romano (1994) in our empirical analysis, we assume that $\mathbf{X}_t = (X_{1,t}, \ldots, X_{l,t})'$ is strictly stationary, $E|\mathbf{X}_t|^{r+\delta} < \infty$ for some $r > 2$ and some $\delta > 0$, and that $\mathbf{X}_t$ is α-mixing of order $-r/(r-2)$. These assumptions are due to Goncalves and de Jong (2003) and are weaker than those formulated in Politis and Romano (1994). The stationarity of $\{\mathbf{X}_t\}$ would be satisfied if $\{r_t\}$ is strictly stationary, because $\{\mathbf{X}_t\}$ is a function of $\{r_t\}$. Next, the moment condition is not alarming, because $\{\mathbf{X}_t\}$ measures the difference in performance of pairs of models, and it is unlikely that the predictions would be so different that the relative loss would violate the moment condition, since the models are quite similar and have the same information. Finally, the mixing condition for $\{\mathbf{X}_t\}$ is satisfied if it holds for r_t. It is important to note that we have not assumed that any of the volatility models are correctly specified. Nor is such an assumption needed, since our ranking of volatility models is entirely measured in terms of expected loss. The assumptions about $\{r_t\}$ will suffice for the comparison and inference, and it is not necessary to make a reference to the true specification of the conditional variance. On the other hand, there is nothing preventing one of the volatility models being correctly specified.[8]

The bootstrap implementation can be justified under weaker assumptions than those above. For example, the stationarity assumption about $\{r_t\}$ can be relaxed and replaced by a near-epoch condition for $\mathbf{X}_t$, see Goncalves and de Jong (2003). This is valuable to have in mind in the present context, since the returns may not satisfy the strict stationarity requirement. A structural change in the DGP would be more critical for our analysis. While a structural change need not invalidate the bootstrap inference (if the break occurs prior to the evaluation period), it would make it very difficult to interpret the results, because the models are estimated using data that have different stochastic properties.

As stated above, the null hypothesis is given by $H_0 : \boldsymbol{\lambda} \leq \mathbf{0}$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_l)'$. The SPA test is based on the test statistic, $T_n^{SPA} \equiv \max_{k=1,\ldots,l} \overline{X}_k / \hat{\omega}_{kk}$, where $\overline{X}_k$ is the kth element of $\overline{\mathbf{X}} \equiv n^{-1} \sum_{t=1}^{n} \mathbf{X}_t$ and $\hat{\omega}_{kk}^2$ is a consistent estimator of $\omega_{kk}^2 \equiv \lim_{n\to\infty} \mathrm{var}(\sqrt{n}\overline{X}_{k,n})$, $k = 1, \ldots, l$. Thus, T_n^{SPA} represents the largest t-statistic (of relative performance) and the relevant question is whether T_n^{SPA} is too large for it to be plausible that $\boldsymbol{\lambda} \leq \mathbf{0}$. This is precisely the question that the test for SPA is designed to answer, as it estimates the distribution of T_n^{SPA} under the null hypothesis and obtains the critical value for T_n^{SPA}.

A closely related test is the RC of White (2000) that employs the non-standardized test statistic $T_n^{RC} \equiv \max_{k=1,\ldots,l} \overline{X}_k$. The critical values of the SPA test and the RC are derived in different ways,

[8] Even the IGARCH model produces a stationary returns series $\{r_t\}$, see Nelson (1990).

and this causes the latter to be sensitive to the inclusion of poor and irrelevant models, and to be less powerful, see Hansen (2003) for details. Power is important for our application, because a more powerful test is more likely to detect superior volatility models, if such exist.

Given the assumptions stated earlier in this section, it holds that $n^{1/2}(\overline{\mathbf{X}} - \boldsymbol{\lambda}) \xrightarrow{d} N_l(\mathbf{0}, \Omega)$, where '$\xrightarrow{d}$' denotes convergence in distribution, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_l)'$ and $\Omega \equiv \lim_{n\to\infty} E[n(\overline{\mathbf{X}} - \boldsymbol{\lambda})(\overline{\mathbf{X}} - \boldsymbol{\lambda})']$. This result makes it possible to test the hypothesis, $H_0 : \boldsymbol{\lambda} \leq \mathbf{0}$.

5.1. Bootstrap Implementation

Unless n is large relative to l it is not possible to obtain a precise estimate of the $l \times l$ covariance matrix, Ω. It is therefore convenient to use a bootstrap implementation, which does not require an explicit estimate of Ω, and the tests of White (2000) and Hansen (2001) can both be implemented with the stationary bootstrap of Politis and Romano (1994).[9] From the bootstrap resamples, $(\mathbf{X}^*_{b,1}, \ldots, \mathbf{X}^*_{b,n})$, $b = 1, \ldots, B$, we can construct random draws of quantities of interest, which can be used to estimate the distributions of these quantities. In our setting we seek an estimate of ω^2_{kk} and estimates of the distributions of T_n^{SPA} and T_n^{RC}. First we calculate the sample averages, $\overline{\mathbf{X}}^*_b \equiv n^{-1}\sum_{t=1}^n \mathbf{X}^*_{b,t}$, $b = 1, \ldots, B$, and it follows from Goncalves and de Jong (2003) that the empirical distribution of $n^{1/2}\overline{\mathbf{X}}^*_b$ converges to the true asymptotic distribution of $n^{1/2}\overline{\mathbf{X}}$. The resamples also allow us to calculate $\hat{\omega}^2_{kk} \equiv \frac{n}{B}\sum_{b=1}^B (\overline{X}^*_{b,k} - \overline{X}_k)^2$, which is consistent for ω^2_{kk}. We seek the distribution of the test statistics, T_n^{SPA} and T_n^{RC}, under the null hypothesis, $\boldsymbol{\lambda} \leq \mathbf{0}$, so we must re-centre the bootstrap variables, such that they satisfy the null hypothesis.[10] Ideally, the variables should be re-centred about the true value of $\boldsymbol{\lambda}$, but since $\boldsymbol{\lambda}$ is unknown we must use an estimate and Hansen (2001) proposed the estimates:

$$\hat{\lambda}^l_k = \min(\overline{X}_k, 0), \quad \hat{\lambda}^c_k = \overline{X}_k 1_{\{\overline{X}_{k,n} \leq -A_{k,n}\}} \quad \text{and} \quad \hat{\lambda}^u_k = 0$$

where $A_{k,n} \equiv \frac{1}{4}n^{-1/4}\hat{\omega}_{kk}$. Thus we define $\overline{Z}^{*,i}_{b,k} = \overline{X}^*_{b,k} - g_i(\overline{X}_k)$, for $i = l, c, u$, where $g_l(\cdot) \equiv \max(x, 0)$, $g_c(x) \equiv x \cdot 1_{\{x > -A_{k,n}\}}$ and $g_u(x) \equiv x$, and it follows that $E[\overline{Z}^{*,i}_{b,k}|\mathbf{X}_1, \ldots, \mathbf{X}_n] = \hat{\lambda}^i_k \leq 0$ for $i = l, c, u$. This enables us to approximate the distribution of T_n^{SPA} by the empirical distribution of

$$T^{SPA*,i}_{b,n} \equiv \max_{k=1,\ldots,l} \frac{n^{1/2}\overline{Z}^{*,i}_{b,k}}{\hat{\omega}_{kk}}, \quad b = 1, \ldots, B, i = l, c, u \tag{2}$$

and we calculate the p-value: $\hat{p}^i_{\text{SPA}} \equiv B^{-1}\sum_{b=1}^B 1_{\{T^{SPA*,i}_{b,n} > T^{SPA}_n\}}$, for $i = l, c, u$. The null hypothesis is rejected for small p-values. In the event that $T_n^{SPA} \leq 0$, there is no evidence against the null hypothesis, and in this case we use the convention: $\hat{p}_{\text{SPA}} \equiv 1$.

The three choices for $\hat{\lambda}_k$ will typically yield three different p-values, and Hansen (2001) has shown that the p-value based on $\hat{\lambda}^c_k$ is consistent for the true p-value, whereas $\hat{\lambda}^l_k$ and $\hat{\lambda}^u_k$ provide an upper and lower bound for the true p-value, respectively.[11] We denote the three resulting tests by SPA_l, SPA_c and SPA_u, where the subscripts refer to lower, consistent and upper. The purpose

[9] This procedure involves a dependence parameter, q, that serves to preserve possible time-dependence in $\mathbf{X}_t$. We used $q = 0.5$ and generated $B = 10{,}000$ bootstrap resamples in our empirical analysis.

[10] The bootstrap variables are constructed such that $E(\overline{\mathbf{X}}^*_b|\mathbf{X}_1, \ldots, \mathbf{X}_n) = \overline{\mathbf{X}}$, and typically we have $\overline{\mathbf{X}} \not\leq \mathbf{0}$.

[11] The true p-value is defined as $\lim_{n\to\infty} P(T_n^{SPA} > t)$, where t is the observed value of the test statistic and the probability is evaluated using the true (but unknown) values of λ and Ω.

of the correction factor, $A_{k,n}$, that defines $\hat{\lambda}_k^c$, is to ensure that $\lim_{n\to\infty} P(\hat{\lambda}_k^c = 0|\lambda_k = 0) = 1$ and $\lim_{n\to\infty} P(\overline{Z}_{b,k,n}^* \leq 0|\lambda_k < 0) = 1$. This is important for the consistency, because the models with $\lambda_k < 0$ do not influence the asymptotic distribution of T_n^{SPA}, see Hansen (2001). However, the choice of $A_{k,n}$ is not unique, and it is therefore useful to include the p-values of the two other tests, SPA_l and SPA_u, because they define the range of p-values that can be obtained by varying the choice for $A_{k,n}$. The p-values based on the tests statistic, T_n^{RC}, are obtained similarly. These are denoted by RC_l, RC_c and RC_u, where RC_u corresponds to the original RC of White (2000).

6. DATA AND EMPIRICAL RESULTS

The models are estimated by maximum likelihood using the estimation sample, and the model's forecasts are compared to the realized variance in the evaluation sample.

The first data set consists of DM–$ spot exchange rate data, where the estimation sample spans the period from October 1, 1987 through September 30, 1992 (1254 observations) and the out-of-sample evaluation sample spans the period from October 1, 1992 through September 30, 1993($n = 260$). The realized variance data for the exchange rate have previously been analysed in Andersen and Bollerslev (1998) and are based on $m = 288$ intraday returns per day. See Andersen and Bollerslev (1997) for additional details. We adjust their measure of realized variance and use $\hat{\sigma}_t^2 \equiv \hat{c} \cdot RV_t^{(288)}$, where $\hat{c} = 0.8418$ is defined in (1).

The second data set consists of IBM stock returns, where the estimation period spans the period from January 2, 1990 through May 28, 1999 (2378 days) and the evaluation period spans the period from June 1, 1999 through May 31, 2000 ($n = 254$). The realized variances were constructed from high-frequency data that were extracted from the Trade and Quote (TAQ) database. The intraday returns, $r_{t,i,m}$, were constructed artificially by fitting a cubic spline to all mid-quotes of a given trading day, using the time interval 9 : 30 EST–16 : 00 EST.[12] From the splines we extract $f = 130$ artificial three-minute returns per day (out of the hypothetical $m = 480$ three-minute returns) and calculate $RV_t^{(130/480)}$. There are several other methods for constructing the realized variance and several of these are discussed in Andersen *et al.* (2003). Later we verify that our empirical results are not influenced by our choice of estimator, as we reach the same conclusions by using six other measures of the realized variance.

The estimate of the adjustment coefficient, (1), is $\hat{c} = 4.4938$, which exceeds $480/130 \simeq 3.7$. This indicates that $RV_t^{(f/m)}$ underestimates the daily variance by more than would be expected if the daily volatility was evenly spread over the 24 hours of the day. There are several possible explanations to the fact that we need to adjust the volatilities by a number different than 3.7. First of all, it could be the result of sample variation, but this seems unlikely as n is too large for sampling error to explain this large a difference. A second explanation is that our intraday returns are positively autocorrelated. The autocorrelation can arise from the market microstructure effects or can be an artifact of the way intraday returns are constructed. A third explanation is that returns are relatively more volatile between close and open, than between open and close, measured per unit of time. This requires that more information arrives to the market while it is closed than while it is open. This contradicts the findings of French and Roll (1986) and Baillie and Bollerslev (1989), so we find this explanation to be unrealistic. Finally, a fourth factor that can

[12] This is done by applying the Splus routine called smooth.spline, which is a one-dimensional cubic smoothing spline that has a basis of B-splines, as discussed in chapters 1–3 of Green and Silverman (1994).

create a difference between squared interday returns and the sum of squared intraday returns is the omission of the conditional expected value $E(r_{t,i,m}|\mathcal{F}_{t-1})$, $i = 1, \ldots, m$ in the calculations. Suppose that $E(r_{t,i,m}|\mathcal{F}_{t-1}) = 0$ for $i = 1, \ldots, f$, but is positive during the time the market is closed. Then r_t^2 would, on average, be larger than $\frac{m}{f}\sum_{i=1}^{f} r_{t,i,m}^2$, even if intraday returns were independent and homoskedastic. Such a difference between expected returns during the time the market is open and closed could be explained as a compensation for the lack of opportunities to hedge against risk overnight. It is not important which of the four explanations cause the difference, as long as our adjustment does not favour some models over others. Because the adjustment is made *ex post* and does not depend on the model forecasts, it is unlikely that a particular model would benefit more than other models.

6.1. Results from the Model Comparison

Table II contains the results from the model comparisons in the form of p-values.[13] The p-values correspond to the hypothesis that the benchmark model, ARCH(1) or GARCH(1,1), is the best model. The naive p-value is the p-value that one would obtain by comparing the best performing model to the benchmark without controlling for the full set of models. So the naive p-value is not a *valid* p-value and it will often be too small, and therefore more likely to indicate an unjustified 'significance'. The p-values of the SPA test and the RC control for the full set of models. Those of SPA_c and RC_c are asymptotically valid p-values, whereas those with subscript l and u provide lower and upper bounds for the p-values. Although the naive p-value is not valid, it can exceed that of the SPA_c, because the best performing model need not be the model that results in the largest t-statistic.

Panel A contains the results for the exchange rate data. The p-values clearly show that the ARCH(1) is outperformed by other models, although the MSE_2 criterion is a possible exception. However, there is no evidence that the GARCH(1,1) is outperformed and a closer inspection of the models reveals that the GARCH(1,1) has one of the best sample performances.

Panels B and C contain the results from the IBM return data, based on the SPA test and the RC, respectively. From Panel B it is evident that both the ARCH(1) and the GARCH(1,l) are significantly outperformed by other volatility models in terms of all loss functions, with the possible exception of the R^2LOG loss function. Thus there is strong evidence that the GARCH(1,1) is inferior to alternative models. The p-values in Panel C are based on the (non-standardized) test statistic T_n^{RC}. The results in Panel C are alarmingly different from those in Panel B, because these p-values suggest the exact opposite conclusion in most cases. Panel C suggests that the GARCH(1,1) is not significantly outperformed, and even the ARCH(1) cannot be rejected as being superior to all other models for three of the six loss functions. The contradicting results are explained by the fact that the T_n^{RC} is not properly standardized, and this causes the tests RC_l, RC_c and RC_u to be sensitive to erratic models. The problem is that a model with a relatively large $\text{var}(\overline{X}_k)$ has a disproportional effect on the distribution of T_n^{RC}, in particular the right tail which defines the critical values, see Hansen (2003). The p-values in the right-most column (boldface) are those of the original RC by White (2000), and these provide little evidence against the two benchmarks. So the results in Table II confirm that the RC is less powerful than the SPA test.

The realized variance can be constructed in many ways and different measures of the realized variance could lead to different results. To verify that our results are not sensitive to our choice

[13] Additional results are given in a technical appendix (Hansen and Lunde, 2001).

Table II. Exchange rate data (DM/USD)

Panel A: Exchange rate data (DM/USD), SPA p-values

	Benchmark: ARCH(1)				Benchmark: GARCH(1,1)			
Metric	Naive	SPA_l	SPA_c	SPA_u	Naive	SPA_l	SPA_c	SPA_u
MSE_1	0.0077	0.0179	**0.0179**	0.0209	0.2911	0.3164	**0.4589**	0.7887
MSE_2	0.0392	0.0695	**0.0748**	0.0797	0.2025	0.6006	**0.7652**	0.9279
QLIKE	0.0067	0.0169	**0.0184**	0.0194	0.2528	0.5831	**0.7707**	0.9639
R^2LOG	<0.0001	0.0002	**0.0002**	0.0002	0.0708	0.2144	**0.3269**	0.6627
MAE_1	<0.0001	0.0002	**0.0002**	0.0002	0.0636	0.2274	**0.3296**	0.6309
MAE_2	0.0002	0.0011	**0.0011**	0.0012	0.1832	0.2177	**0.2920**	0.5663

Panel B: IBM Data, SPA p-values

	Benchmark: ARCH(1)				Benchmark: GARCH(1,1)			
Metric	Naive	SPA_l	SPA_c	SPA_u	Naive	SPA_l	SPA_c	SPA_u
MSE_1	0.0052	0.0002	**0.0002**	0.0002	0.0355	0.0245	**0.0300**	0.0358
MSE_2	0.0061	0.0001	**0.0001**	0.0001	0.0409	0.0260	**0.0288**	0.0316
QLIKE	0.0003	<0.0001	<**0.0001**	<0.0001	0.0213	0.0379	**0.0463**	0.0528
R^2LOG	0.0108	0.0011	**0.0011**	0.0014	0.0166	0.0526	**0.0630**	0.0741
MAE_1	0.0012	0.0080	**0.0086**	0.0104	0.0026	0.0040	**0.0051**	0.0058
MAE_2	0.0014	0.0097	**0.0100**	0.0115	0.0026	0.0054	**0.0065**	0.0078

Panel C: IBM Data, RC p-values

	Benchmark: ARCH(1)				Benchmark: GARCH(1,1)			
Metric	Naive	RC_l	RC_c	RC_u	Naive	RC_l	RC_c	RC_u
MSE_1	0.0052	0.0164	0.0164	**0.0164**	0.0355	0.1000	0.1499	**0.2811**
MSE_2	0.0061	0.0205	0.0205	**0.0205**	0.0409	0.1053	0.1056	**0.1472**
QLIKE	0.0003	0.0017	0.0017	**0.0017**	0.0213	0.0943	0.1153	**0.3750**
R^2LOG	0.0108	0.0601	0.0713	**0.0713**	0.0166	0.2908	0.3535	**0.6039**
MAE_1	0.0012	0.0972	0.1227	**0.1399**	0.0026	0.0505	0.1144	**0.1522**
MAE_2	0.0014	0.1219	0.1649	**0.1941**	0.0026	0.0644	0.1135	**0.1734**

Notes: The table presents p-values of the SPA test and the RC for two null hypotheses: that the benchmark model, ARCH(1) or GARCH(1,1), is the best model. Conclusions should be based on the SPA_c test (boldface) in Panels A and B. The naive 'p-value' compares the best performing model to the benchmark, but ignores the full set of models. So the naive 'p-value' is not a valid p-value and the difference between it and that of SPA_c(RC_c) shows the effects of data mining. Panel C contains the p-values that are based on the RC non-standardized test statistic. The p-values of the original RC are in boldface. A comparison of the results of Panels B and C shows that the SPA test is more powerful than the RC, and the latter is unable to detect the inferiority of the GARCH(1,1), and the ARCH(1) in some cases.

of RV measure we repeat the empirical analysis of the IBM returns data using six other measures. These measures include: one based on a different spline method and sampling frequency; one based on the Fourier method by Barucci and Reno (2002); two based on the previous-tick method; and two based on the linear interpolation method. The p-values of the SPA_c test for the seven different measures of the realized variance are presented in Table III. Fortunately, the p-values do not differ much across the various measures of the realized variance, although most of the alternative measures provide slightly stronger evidence that the GARCH(1,1) is outperformed in terms of the R^2LOG loss function, and slightly weaker evidence in terms of the MAE_1 and MAE_2 loss functions.

Table III. Results for different measures of realized variance

Criterion	Method for estimating realized variance						
	Spl-50 3 min	Spl-250 2 min	Fourier $M = 85$	Linear 5 min	Previous 5 min	Linear 1 min	Previous 1 min
MSE_1	0.0271	0.0230	0.0134	0.0125	0.0133	0.0111	0.0103
MSE_2	0.0280	0.0213	0.0135	0.0168	0.0181	0.0082	0.0082
QLIKE	0.0457	0.0350	0.0166	0.0178	0.0175	0.0112	0.0118
R^2LOG	0.0651	0.0998	0.0462	0.0409	0.0505	0.0375	0.0340
MAE_1	0.0039	0.0635	0.0476	0.0690	0.0662	0.0960	0.0881
MAE_2	0.0056	0.0888	0.0724	0.0510	0.0600	0.0707	0.0749

Notes: This table reports p-values of the SPA_c test from the analysis of IBM returns where the GARCH(1,1) is used as the benchmark. The p-values are obtained for seven different measures of the realized variance that are constructed with different techniques (and sampling frequencies). Spl-50 and Spl-250 refer to a cubic spline method that use 50 and 250 knot points, respectively; the third measure is based on the Fourier method; and the last four measures are based on the linear interpolation and previous-tick methods.

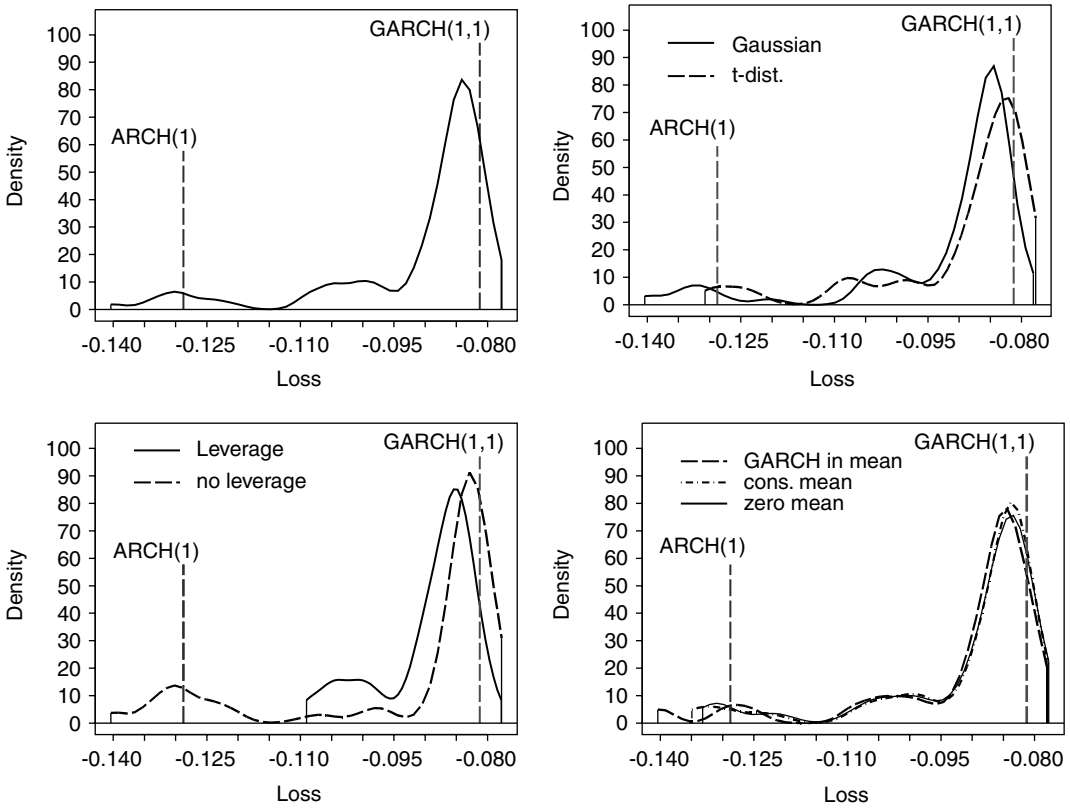

Figure 1. Population of model performance: exchange rate data and MSE_2 loss function. The x-axis is the negative value of average sample loss

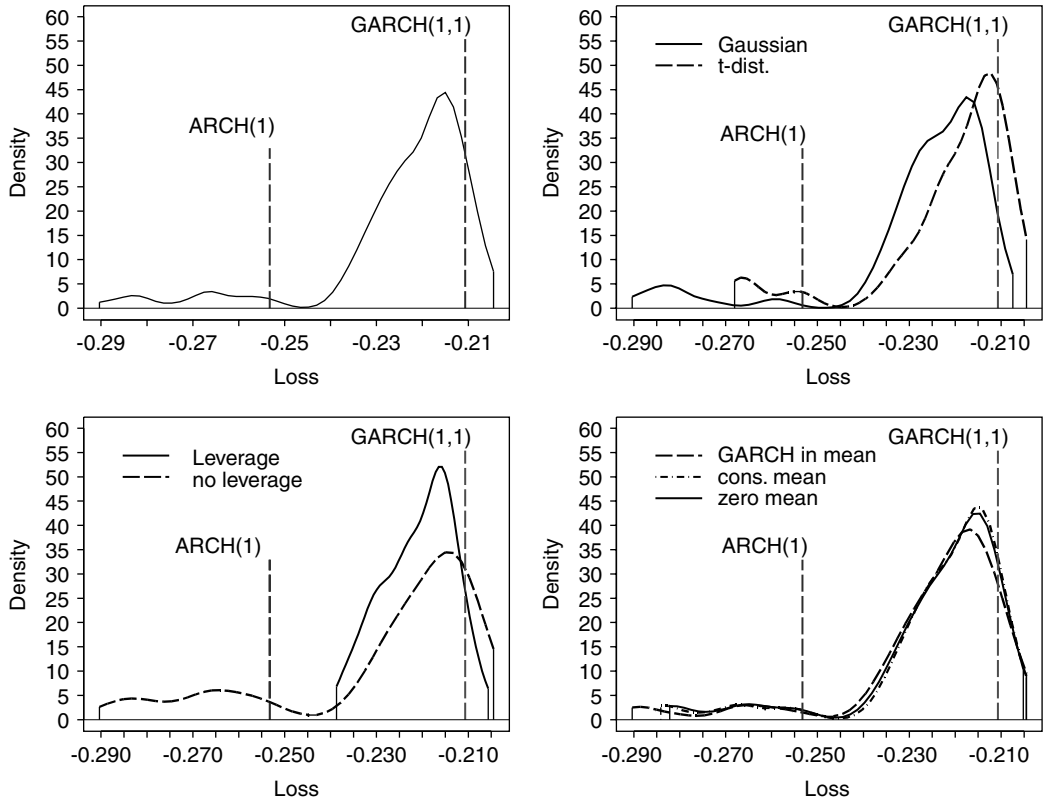

Figure 2. Population of model performance: exchange rate data and MAE_2 loss function. The x-axis is the negative value of average sample loss

Figures 1–4 show the 'population' of model performances for various loss functions (and the two data sets).[14] The plots provide information about how similar/different the models' sample performances were, and show the location of the ARCH(1) and GARCH(1,1) relative to the full set of models. The x-axis is the (negative value of) average sample loss, such that the right tail represents the model with the best sample performance. Each figure contains four panels. The upper left panel is the model density of all the models, whereas the last three panels show the performance densities for different 'types' of models. The models are divided into groups according to their type: Gaussian vs. t-distributed specification; models with and without a leverage effect; and the three mean specifications.

Figures 1 and 2, which display the results for the exchange rate data, show that the GARCH(1,1) is one of the best performing models, whereas the ARCH(1) has one of the worst sample performances. There are no major differences between the various types of models, although there is a small tendency that the t-distributed specification leads to a better performance than a Gaussian specification in Figure 2.

[14] To save space, we have only included the figures that correspond to the MSE_2 and MAE_2 loss functions. The figures for all six loss functions are given in Hansen and Lunde (2001).

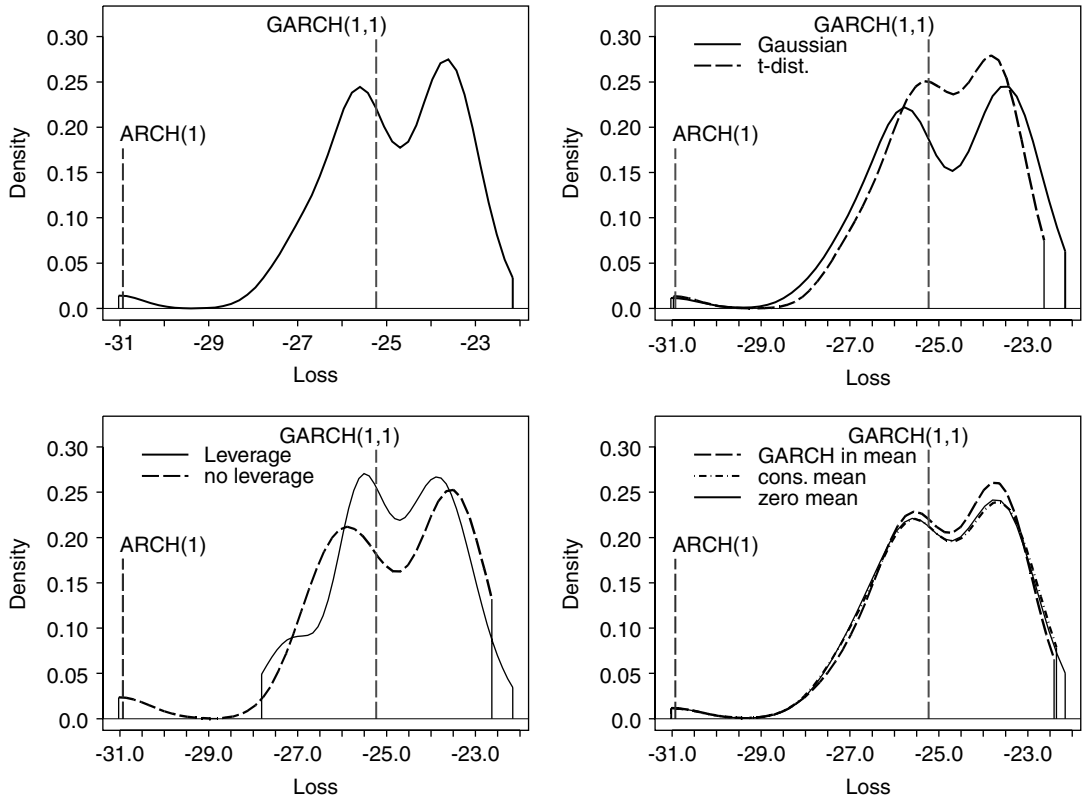

Figure 3. Population of model performance: IBM data and MSE_2 loss function. The x-axis is the negative value of average sample loss

The results for the IBM return data are displayed in Figures 3 and 4. From the SPA test we concluded that the GARCH(1,1) was significantly outperformed by other models, and the two figures also show that the GARCH(1,1) is ranked much lower in this sample. It now seems that the Gaussian specification does better than the t-distributed specification, on average. However, the very best performing model in terms of the MAE_2 loss function is a model with a t-distributed specification. From our analysis of the IBM data it is evident that models that can accommodate a leverage effect are superior to those that cannot, particularly in Figure 4.

Although the conditional mean $\mu_t = E(r_t|\mathcal{F}_{t-1})$ is likely to be small, it cannot *ex ante* be ruled out that a more sophisticated specification for μ_t, such as the GARCH-in-mean, leads to better forecasts of volatility than the zero-mean specification. However, the performance is almost identical across the three mean specifications, as can be seen from Figures 1–4.

7. CONCLUSIONS

We have compared a large number of volatility models, in terms of their ability to forecast the conditional variance in an out-of-sample setting.

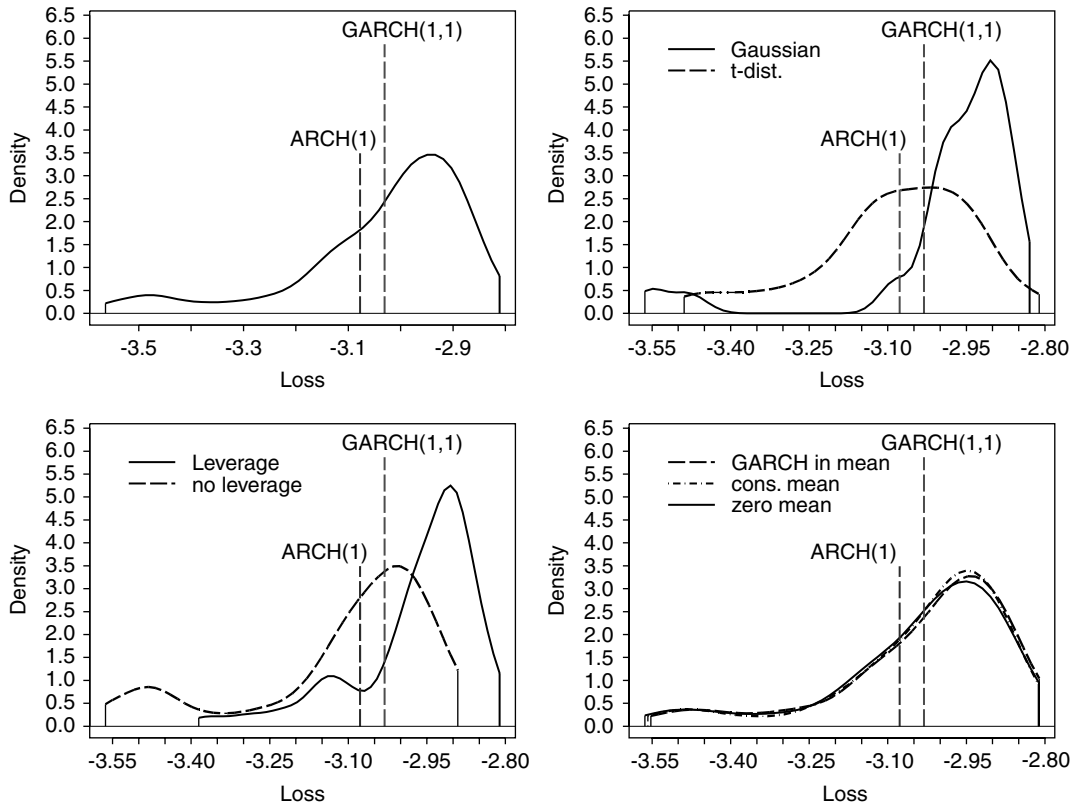

Figure 4. Population of model performance: IBM data and MAE_2 loss function. The x-axis is the negative value of average sample loss

Our analysis was limited to DM–$ exchange rates and IBM stock returns and a universe of models that consisted of 330 different ARCH-type models. The main findings are that there is no evidence that the GARCH(1,1) model is outperformed by other models, when the models are evaluated using the exchange rate data. This cannot be explained by the SPA test lacking power because the ARCH(1) model is clearly rejected and found to be inferior to other models. In the analysis of IBM stock returns we found conclusive evidence that the GARCH(1,1) is inferior, and our results strongly suggested that good out-of-sample performance requires a specification that can accommodate a leverage effect.

The performances of the volatility models were measured out-of-sample using six loss functions, where realized variance was used to construct an estimate of the unobserved conditional variance. The significance of relative performance was evaluated with the test for superior predictive ability of Hansen (2001) and the reality check for data snooping of White (2000). Our empirical analysis illustrated the usefulness of the SPA test and showed that the SPA test is more powerful than the RC.

The SPA test and the RC are not model selection criteria and therefore not designed to identify the best volatility model (in population). It is also unlikely that our data contain sufficient information to conclude that the model with the best sample performance is significantly better

than all other models. Nevertheless, the use of a significance test, such as the SPA test, has clear advantages over model selection criteria, because it allows us to make strong conclusions. In our setting, the SPA test provided conclusive evidence that the GARCH(1,1) is inferior to other models in our analysis of IBM returns. However, in the analysis of the exchange rate data, there was no evidence against the claim that: 'nothing beats a GARCH(1,1)'.

ACKNOWLEDGEMENTS

Financial support from the Danish Research Agency, grant no 24-00-0363, and the Salomon Research Award at Brown University is gratefully acknowledged. We thank Professor M. Hashem Pesaran (editor) and two anonymous referees for many suggestions that improved our paper, and we thank Tim Bollerslev for sharing the realized variance data for the exchange rate and Roberto Renò for constructing some of the realized variance data for the IBM returns. We also thank Tim Bollerslev, Frank Diebold, Rob Engle and seminar participants at Aarhus School of Business, Aarhus University, Brown University, Cornell University, University of Pennsylvania and ESEM 2002 for valuable comments. We are responsible for all remaining errors.

REFERENCES

Andersen TG, Bollerslev T. 1997. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance* **4**: 115–158.

Andersen TG, Bollerslev T. 1998. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review* **39**(4): 885–905.

Andersen TG, Bollerslev T, Diebold FX. 2003. Parametric and nonparametric volatility measurement. In *Handbook of Financial Econometrics*, Vol. I, Aït-Sahalia Y, Hansen LP (eds). Elsevier-North Holland: Amsterdam.

Baillie RT, Bollerslev T. 1989. The message in daily exchange rates: a conditional variance tale. *Journal of Business & Economic Statistics* **7**(4): 297–305.

Barndorff-Nielsen OE, Shephard N. 2001. Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics (with discussion). *Journal of the Royal Statistical Society, Series B* **63**(2): 167–241.

Barucci E, Reno R. 2002. On measuring volatility of diffusion processes with high frequency data. *Economics Letters* **74**: 371–378.

Black F. 1976. Studies in stock price volatility changes. Proceedings of the 1976 Business Meeting of the Business and Economics Section, American Statistical Association; 177–181.

Bollerslev T. 1986. Generalized autoregressive heteroskedasticity. *Journal of Econometrics* **31**: 307–327.

Bollerslev T, Engle RF, Nelson D. 1994. ARCH models. In *Handbook of Econometrics*, Vol. IV, Engle RF, McFadden DL (eds). Elsevier Science B.V.: Amsterdam; 2961–3038.

Diebold FX, Lopez JA. 1996. Forecast evaluation and combination. In *Handbook of Statistics*, Vol. 14, Statistical Methods in Finance, Maddala GS, Rao CR (eds). North-Holland: Amsterdam; 241–268.

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–263.

Ding Z, Granger CWJ, Engle RF. 1993. A long memory property of stock market returns and a new model. *Journal of Empirical Finance* **1**: 83–106.

Duan J. 1997. Augmented GARCH(p,q) process and its diffusion limit. *Journal of Econometrics* **79**(1): 97–127.

Engle RF. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* **45**: 987–1007.

Engle RF, Ng V. 1993. Measuring and testing the impact of news on volatility. *Journal of Finance* **48**: 1747–1778.

Engle RF, Patton AJ. 2001. What good is a volatility model? *Quantitative Finance* **1**(2): 237–245.
Fleming J, Kirby C, Ostdiek B. 2003. The economic value of volatility timing using realised volatility. *Journal of Financial Economics* **67**: 473–509.
French KR, Roll R. 1986. Stock return variance: the arrival of information and the reaction of traders. *Journal of Financial Economics* **17**: 5–26.
Goncalves S, de Jong R. 2003. Consistency of the stationary bootstrap under weak moment conditions. *Economics Letters* **81**: 273–278.
Green PJ, Silverman BW. 1994. *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall: London.
Hansen BE. 1994. Autoregressive conditional density models. *International Economic Review* **35**(3): 705–730.
Hansen PR. 2001. A test for superior predictive ability. Brown University, Department of Economics Working Paper 2001–06 (http://www.econ.brown.edu/fac/Peter_Hansen).
Hansen PR. 2003. Asymptotic tests of composite hypotheses. Brown University, Department of Economics Working Paper 2003–09 (http://www.econ.brown.edu/fac/Peter_Hansen).
Hansen PR, Lunde A. 2001. Consistent ranking of volatility models (http://www.hha.dk/∼alunde/academic/research/papers/vola-mod-appendix.pdf).
Hansen PR, Lunde A. 2003. Consistent preordering with an estimated criterion function, with an application to the evaluation and comparison of volatility models. Brown University Working Paper 2003–01 (http://www.econ.brown.edu/fac/Peter_Hansen).
Hansen PR, Lunde A, Nason JM. 2003. Choosing the best volatility models: the model confidence set approach. *Oxford Bulletin of Economics and Statistics* **65**: 839–861.
Hentshel L. 1995. All in the family: nesting symmetric and asymmetric garch models. *Journal of Financial Economics* **39**: 71–104.
Hol E, Koopman SJ. 2002. Stock index volatility forecasting with high frequency data. Manuscript, Department of Econometrics, Free University of Amsterdam.
Lopez JA. 2001. Evaluation of predictive accuracy of volatility models. *Journal of Forecasting* **20**(1): 87–109.
Loudon GF, Watt WH, Yadav PK. 2000. An empirical analysis of alternative parametric ARCH models. *Journal of Applied Econometrics* **15**(1): 117–136.
Martens M. 2002. Measuring and forecasting S&P 500 index futures volatility using high-frequency data. *Journal of Futures Markets* **22**(6): 497–518.
Meddahi N. 2002. A theoretical comparison between integrated and realized volatility. *Journal of Applied Econometrics* **17**: 479–508.
Nelson DB. 1990. Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory* **6**: 318–334.
Pagan AR, Schwert GW. 1990. Alternative models for conditional volatility. *Journal of Econometrics* **45**: 267–290.
Politis DN, Romano JP. 1994. The stationary bootstrap. *Journal of the American Statistical Association* **89**: 1303–1313.
Poon S-H, Granger C. 2003. Forecasting volatility in financial markets: a review. *Journal of Economic Literature* **41**: 478–539.
Schwert GW. 1990. Stock volatility and the crash of '87. *Review of Financial Studies* **3**(1): 77–102.
Taylor SJ. 1986. *Modelling Financial Time Series*. John Wiley & Sons: New York.
West KD. 1996. Asymptotic inference about predictive ability. *Econometrica* **64**: 1067–1084.
White H. 2000. A reality check for data snooping. *Econometrica* **68**: 1097–1126.

