

A Forecast Comparison of Volatility Models:
Does Anything Beat a GARCH(1, 1)?

Peter Reinhard Hansen¹

and

Asger Lunde²

Working Paper No. 01-04

First Draft, March, 2001

Revised, November, 2001



Brown University

Department of Economics

¹Brown University, Department of Economics, Box B, Brown University, Providence, RI 02912, USA, Phone: (401) 863 9864, Email: Peter_Hansen@brown.edu

²The Aarhus School of Business, Department of Information Science, Fuglesangs Allé 4 DK-8210 Aarhus V, Phone (+45) 89486688, Email: alunde@asb.dk

Abstract³

We compare a large number of volatility models in terms of their ability to describe the behavior of the behavior of conditional variance, using out-of-sample data. Our question of interest is whether more sophisticated volatility models are able to outperform the simple GARCH(1, 1) model. This question is addressed using the test for superior predictive ability (SPA) by Hansen (2001). A salient property of this test is that it takes the performance of all models into account simultaneously, thereby avoiding crude approximations and the distortion that arises from pair-wise comparisons.

When the models are compared using DM-\$ exchange rate data, we do not find evidence that the GARCH(1, 1) is outperformed by other models. However, when we compare the models using IBM equity return data, we find the GARCH(1, 1) to be significantly outperformed by alternative models. Most of the models that perform well in this data set are models that can accommodate a leverage effect.

Our analysis confirms that the test for SPA of Hansen (2001) is more powerful than the Reality Check (RC) of White (2000). The RC is unable to reject the GARCH(1, 1) as a superior model in both data sets and, in fact, the RC is in most cases unable to find evidence that a simple ARCH(1) model is outperformed by other models. The SPA test always finds the ARCH(1) to be an inferior forecasting model.

JEL Classification: C12; C13; C15; C22; C52; C53; G15

Keywords: Volatility Models, Forecast Comparison, Superior Predictive Ability

³We thank: Tim Bollerslev for valuable comments and for making the exchange rate data available to us, Roberto Renò for construction several intra-day measures of conditional volatility for our IBM data, and we thank Sivan Ritz for suggesting numerous clarifications. All errors remain our responsibility.

1 Introduction

The conditional variance of financial time-series is important when pricing derivatives, calculating measures of risk, and hedging against portfolio risk. Therefore, there has been an enormous interest amongst researchers and practitioners to model the conditional variance. As a result, a large number of volatility models have been developed since the seminal paper of Engle (1982).

The aim of this paper is to examine whether more sophisticated volatility models provide a better description of financial time-series than parsimonious models. In order to answer this question, we need to address the following four intermediate questions: (1) Which models should be included in the comparison; (2) How should the volatility models be evaluated; (3) How to handle the fact that the conditional variance is unobserved; and (4) How to make inference in a comparison of multiple (non-nested) models.

A brief description of the approach we take in this paper is the following: We estimate and compare 330 GARCH-type models in terms of their ability to describe the conditional variance, using the out-of-sample methodology. We use intra-day data to obtain a precise estimate of the conditional variance, often called *realized volatility*, and this estimate is substituted for the unobserved conditional variance. The various volatility models are compared using the test for superior predictive ability of Hansen (2001).

We apply this framework using two data sets. The first data set contains returns (changes) in DM-\$ exchange rate and the second contains returns on the IBM stock. In our analysis of exchange rate data we find no evidence that the GARCH(1, 1) is outperformed. However, in the analysis of the IBM data the GARCH(1, 1) is clearly outperformed – primarily by models that can accommodate a leverage effect. A comparison of a Gaussian versus a t -distributed specification of standardized returns, shows that the t -specification does (on average) better than the Gaussian in the analysis of exchange rates, whereas the opposite is the case in our analysis of IBM returns. The performances of the different mean specifications, zero-mean, constant mean, or GARCH-in-mean, are almost identical.

With reference to the four questions listed above, we proceed as follows.

1. To maintain a manageable scope of the paper, we confine the attention to GARCH-type models. Although we estimate and compare a total of 330 models, we have by no means

exhausted the universe of volatility models that have been proposed in the literature. For example, our analysis does not include stochastic volatility models, FIGARCH, or the recent VAR models of realized volatility by Andersen, Bollerslev, Diebold & Labys (2001). Further, this paper does not address whether continuous-time models provide a better or worse description of volatility than discrete-time models. Nevertheless, our model-space contains models with many distinct characteristics, which are interesting to compare. For example, some models allow the volatility to react asymmetrically to positive and negative changes in returns, known as the leverage effect. Features of this kind are typically found to be very significant in in-sample analyses, but a model that accommodates a significant (in-sample) relation need not result in better out-of-sample performance than that of a simpler model.¹ Other distinct characteristics involve the choice of lag-length, the specification of the mean, and the distributional assumptions for the standardized returns.

2. Ideally, a comparison of volatility models should involve a comprehensive comparison of the models ability to describe all aspects of the conditional distribution. However, a large number of observations are needed to get a good assessment of a distribution, and even more observations may be needed in order to rank the accuracy of multiple predictive distributions. A more powerful comparison can be made if one restricts the comparison to a particular property of the distribution. In our analysis, we compare the volatility models in terms of their ability to describe the daily volatility. In our opinion, this is a natural metric for comparing volatility models, since the main component of a volatility model is an equation that describes the behavior of the conditional variance – typically one-period-ahead.

A complete description of the evaluation requires a specification of a loss function. As was pointed out by Bollerslev, Engle & Nelson (1994), Diebold & Lopez (1996), and Lopez (2001), it is not obvious which loss function should be employed. Given the lack of a unique criterion, we employ six different criteria in our comparison. These include standard criteria such as the mean squared error (MSE) criterion, a likelihood criterion,

¹The reason is that uncertainty from parameter estimation can distort the forecasts more than the omission of explanatory variable. Thus, a misspecified model may yield better forecasts than a correctly specified model.

and the mean absolute deviation criterion.

3. The evaluation and comparison of volatility models is made difficult by the fact that the conditional variance is unobservable. This has made it difficult to identify poor models, and may explain that so many volatility models have been able to coexist. The first approach to circumvent this problem, was to substitute squared returns for the unobserved conditional variance. But this commonly led to a very poor out-of-sample performance, which instigated a discussion of the practical relevance of these models. However, this skepticism was refuted by Andersen & Bollerslev (1998*a*). Rather than using squared inter-day returns, which are very noisy measures of daily volatility, Andersen and Bollerslev based their evaluation on realized volatility, which is an estimate of the volatility calculated from squared intra-day returns. The use of realized volatility revealed that volatility models have a good out-of-sample performance, and that the previously found poor performance could be explained by the use of a noisy measure of the volatility.

Another important argument for using an intra-day estimate of daily volatility, is that it makes it easier to tell good and bad volatility models apart. In fact, a noisy estimate of daily volatility can severely distort the comparison, and may result in an inconsistent comparison. Our comparison is therefore based on an estimate of the conditional variance that are extracted from intra-day data.

In the analysis of the IBM data we face an additional complication, due to limited availability of intra-day data. High frequent intra-day data are only available during the time where the market is open, and a simple estimate of the daily volatility is therefore not available. Nevertheless, under fairly weak assumptions we can make an adjustment that allows us to extract a good estimate of the volatility. This adjustment have an additional advantage, as it automatically removes a potential bias that can arise from the construction of intra-day returns and market micro-structures.

4. Comparing multiple models is a non-standard problem. One of the complications that arise when comparing several models, is that spurious results may appear. A decent, but not superior model can be “lucky” in a particular sample, and appear to be better than all other models. The more models that are being compared, the higher the probability

that some model is going to appear superior by chance. So when comparing models, it is important to take all the models, which have been evaluated, into account along with the interdependence of model performances, across all models.

Major contributions to this problem have been made by Diebold & Mariano (1995) and West (1996), and more recently by White (2000) along with the refinements of Hansen (2001). In this paper we apply the test for superior predictive ability (SPA) of Hansen (2001) to compare the volatility models. This test makes it possible to test whether a particular model (benchmark model) is significantly outperformed by other models, while taking into account the large number of models that are being compared. The test controls for the “mining” over models and can evaluate whether an observed model performance could have occurred by chance, or is sufficiently good to conclude that the model is superior. The approach to inference of the SPA test is very different from tests based on the Bonferroni bound, as the former avoids the conservative approximation that the latter leads to.

This paper is organized as follows. Section 2 describes the 330 volatility models we have estimated and compared. Section 3 describes the loss functions that we employ in the comparison. In Section 4 we describe the intra-day estimation of the conditional variance and Section 5 describes the test for SPA as well as the bootstrap implementation of the test. Our results are presented in Section 6, and Section 7 contains concluding remarks.

2 The GARCH Universe

We use the notation of Hansen (1994) to characterize our universe of parametric GARCH models. In this setting the aim is to model the distribution of some stochastic variable, r_t , conditional on some information set, \mathcal{F}_{t-1} . Formally, \mathcal{F}_{t-1} is the σ -algebra induced by all variables that are observed at time $t - 1$. Thus, \mathcal{F}_{t-1} contains the lagged values of r_t and other predetermined variables.

The variables of interest in our analysis are returns defined from daily prices, p_t . We define the compounded return by

$$r_t = \log(p_t) - \log(p_{t-1}), \quad t = -R + 1, \dots, n, \quad (1)$$

which is the return from holding the asset from time $t - 1$ to time t . The sample period consists of an estimation period with R observations, $t = -R + 1, \dots, 0$, and an evaluation period with n periods, $t = 1, \dots, n$.

The objective is to model the conditional density of r_t , denoted by

$$f(r|\mathcal{F}_{t-1}) \equiv \frac{d}{dr} P(r_t \leq r|\mathcal{F}_{t-1}).$$

In the modelling of the conditional density it is convenient to define the conditional mean, $\mu_t \equiv E(r_t|\mathcal{F}_{t-1})$, and the conditional variance, $\sigma_t^2 \equiv \text{var}(r_t|\mathcal{F}_{t-1})$ (assuming that they exist). Subsequently we can define the standardized returns, which are denoted by $e_t = (r_t - \mu_t)/\sigma_t$, $t = -R + 1, \dots, n$. We denote the conditional density function of the standardized returns by $g(e|\mathcal{F}_{t-1}) = \frac{d}{de} P(e_t \leq e|\mathcal{F}_{t-1})$, and it is simple to verify that the conditional density of r_t is related to the conditional density of e_t , since

$$f(r|\mathcal{F}_{t-1}) = \frac{1}{\sigma_t} g(e|\mathcal{F}_{t-1}).$$

Thus, the modelling of the conditional distribution of r_t can be divided into three elements: the conditional mean, the conditional variance, and the density function of the standardized residuals. This makes the modelling more tractable and makes it easier to interpret a particular specification. In our modelling, we choose a parametric form of the conditional density, starting with the generic specification

$$f(r|\psi(\mathcal{F}_{t-1}; \theta)),$$

where θ is a finite-dimensional parameter vector, and $\psi_t = \psi(\mathcal{F}_{t-1}; \theta)$ is a *time-varying* parameter vector of low dimension. Given a value of θ , we require that ψ_t is observable² at time $t - 1$. This yields a complete specification of the conditional distribution of r_t .

We divide the vector of time-varying parameters into three components,

$$\psi_t = (\mu_t, \sigma_t^2, \eta_t),$$

where μ_t is the conditional mean (the *location* parameter), σ_t is the conditional standard deviation (the *scale* parameter), and η_t are the remaining (*shape*) parameters of the conditional

²This assumption excludes the class of stochastic volatility models from the analysis.

distribution. Hence, our family of density functions for r_t is a location-scale family with (possibly time-varying) shape parameters.

Our notation for the modelling of the conditional mean, μ_t , is given by

$$m_t = \mu(\mathcal{F}_{t-1}; \theta).$$

The conditional mean, μ_t , is typically of secondary importance for GARCH-type models. The primary objective is the conditional variance, σ_t^2 , which is modelled by

$$h_t^2 = \sigma^2(\mathcal{F}_{t-1}; \theta). \tag{2}$$

A complete specification is achieved through a modelling of the density function for the standardized residuals, e_t , which we denote by $g(e|\eta_t)$, where η_t contains the shape parameters. Most of the existing GARCH-type models can be expressed in this framework, and the corresponding η_t 's are typically constant. For example, the earliest models assumed the density $g(e|\eta_t)$ to be (standard) Gaussian. In our analysis we also keep η_t constant. Models with non-constant η_t include Hansen (1994) and Harvey & Siddique (1999). Further, as pointed out by Tauchen (2001), it is possible to avoid restrictive assumptions, and estimate a time-varying density for e_t by semi-nonparametric (SNP) techniques, see Gallant & Tauchen (1989).

2.1 The Conditional Mean

Our modelling of the conditional mean, μ_t , takes the form

$$m_t = \mu_0 + \mu_1 \sigma_{t-1}^2.$$

The three specifications we include in the analysis are: the GARCH-in-mean suggested by Engle, Lilien & Robins (1987), the constant mean ($\mu_1 = 0$), and the zero-mean model ($\mu_0 = \mu_1 = 0$), advocated by Figlewski (1997), see Table 1 for details.

2.2 The Conditional Variance

The conditional variance is the main object of interest, and our analysis includes a large number of parametric specifications for σ_t . These include the ARCH model by Engle (1982), the GARCH model by Bollerslev (1986), the IGARCH model, the Taylor (1986)/Schwert (1989)

(TS-GARCH) model, the A-GARCH³, the NA-GARCH and the V-GARCH models suggested by Engle & Ng (1993), the threshold GARCH model (Thr.-GARCH) by Zakoian (1994), the GJR-GARCH model of Glosten, Jagannathan & Runkle (1993), the log-ARCH by Geweke (1986) and Pantula (1986), the EGARCH of Nelson (1991), the NGARCH of Higgins & Bera (1992), the A-PARCH model proposed in Ding, Granger & Engle (1993), the GQ-ARCH suggested by Sentana (1995), the H-GARCH of Hentshel (1995), and finally the Aug-GARCH suggested by Duan (1997). See Table 2.

Several of the models nest other models as special cases. In particular the H-GARCH and the Aug-GARCH specifications are very flexible specifications of the volatility, and both specifications include several of the other models as special cases.

Not all of these models have are frequently seen in applied work. For example, we do not know of published work that has applied the Aug-GARCH model. Nevertheless, we include all models in our analysis, because of the fact that applications of a particular model have not appeared in published work, does not disqualify it from being relevant for our analysis. The reason is that we seek to get a precise assessment of how good a performance (or excess performance) one can expect to achieve by chance, when estimating a large number of models. Therefore, it is important that we include as many of the existing models as possible, and not just those that were successful in an applied sense and therefore appear in published work. Although, this leads to a very large number of different volatility models, we have by no means exhausted the space of possible GARCH-type model.

The evolution of volatility models has been motivated by empirical findings and economic interpretations. Ding et al. (1993) demonstrated with Monte-Carlo studies that both the original GARCH model by Bollerslev (1986) and the GARCH model in standard deviations, attributed to Taylor (1986) and Schwert (1990), are capable of producing the pattern of autocorrelation that appears in financial data. So in this respect there is not an 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. This is the motivation for the introduction of the *Box-Cox transformation*

³At least four authors have adopted the acronym A-GARCH for different models. To undo this confusion we reserve the A-GARCH name for a model by Engle & Ng (1993) and rename the other models, e.g., the model by Hentshel (1995) is here called H-GARCH.

of the conditional standard deviation and the asymmetric absolute residuals. The observed *leverage effect* motivated the development of models that allowed for an asymmetric response in volatility to positive and negative shocks. The leverage effect was first noted in Black (1976), and it refers to a negatively correlated between returns and changes in the volatility. This implies that volatility should tend to rise in response to bad news, (defined as returns that are lower than expected), and should tend to fall after good news. Given a particular volatility model, one can plot σ_t^2 against ε_{t-1} , which illustrates how the volatility reacts to the difference between realized return and expected return. This plot is a simple way to characterize some of the differences there are among the various specifications of volatility. This method was introduced by Pagan & Schwert (1990), and later named the *News Impact Curve* by Engle & Ng (1993). The News Impact Curve provides an easy way to interpret some aspects of the different volatility specifications. Several of the models included in our analysis were compared using this method by Hentshel (1995).

The specifications for the conditional variance, given in Table 2, contain parameters for the lag lengths, denoted by p and q . We have included the four combinations of lag lengths $p, q = 1, 2$ for all models, with the exceptions of the ARCH, H-GARCH, and Aug-GARCH models. For the H-GARCH, and Aug-GARCH models we only include the specification with $(p, q) = (1, 1)$, as they are very burdensome to estimate. The only ARCH model we have included is the ARCH(1) model, (which corresponds to $(p, q) = (1, 0)$). It is well known that a ARCH model with relatively few lags is unable to capture the persistence in volatility. The ARCH(1) model is only included in our analysis as a point of reference, and to verify that our model comparison is powerful enough to distinguish between models. Because we expect the ARCH(1) to be a poor model, we want to verify that the test for SPA is able to reject it as a superior model. If the test is unable to reject a poor model, then the test is not very informative about the quality of models in the particular sample that is being investigated. It is clearly a restriction, that we only include models with two lags or less. Nevertheless, if a particular specification with two lags is unable to outperform a simple benchmark model, then there is little reason to expect more lags will result in superior forecasts.

2.3 The Density for the Standardized Returns

We only consider a Gaussian and a t -distributed specification for the density $g(e|\boldsymbol{\eta}_t)$; the latter was first advocated by Bollerslev (1987). Thus, $\boldsymbol{\eta}_t$ is held constant in our analysis. In fact, the Gaussian specification is free of parameters, whereas the t -specification has the degrees of freedom as the parameter. Thus, in the estimation of the models with a t -specification, we estimated the degrees of freedom, $\boldsymbol{\eta}$, (which is not restricted to be an integer).

3 Forecast Evaluation

It is non-trivial to evaluate the quality of a volatility model. As pointed out by Bollerslev et al. (1994), there is not a unique criterion for selecting the best model; rather it will depend on preferences, e.g., expressed in terms of a utility function or a loss function. The standard model selection criteria of Akaike and Schwartz are often applied, but this approach is problematic whenever the distributional assumptions (that underlies the likelihood) are dubious. Also, as advertisements for mutual funds always remind us: good in-sample performance does not guarantee good out-of-sample performance. This point is clearly relevant for our analysis. Most of the models we estimate have significant lags in our in-sample analysis (that is p or $q = 2$). But in the out-of-sample comparison, the models with more lags rarely perform better than the corresponding models with fewer lags.

Following the notation of White (2000), we index the l volatility models by k , and denote model k 's forecast of σ_t^2 by $h_{k,t}^2$, $k = 1, \dots, 330$ and $t = 1, \dots, n$. The volatility models ability to make accurate predictions of the volatility, have often been measured in terms of the R^2 from the regression of squared returns on the volatility forecast, that is

$$r_t^2 = a + bh_t^2 + u_t. \tag{3}$$

Unfortunately, this regression is sensitive to extreme values of r_t^2 , especially if estimated by least squares. This implies that the parameter estimates of a and b will primarily be determined by the observations where the squared returns, r_t^2 , have the largest values. This has been noted

by Pagan & Schwert (1990) and Engle & Patton (2000)⁴. Therefore they advocate the regression

$$\log(r_t^2) = a + b \log(h_t^2) + u_t \quad (4)$$

which is less sensitive to “outliers”, because severe mispredictions are given less weight than is the case in (3).

In our analysis, we compare the models in terms of loss functions, some of which are more robust to outliers, than is the regression (4), and we argue that the use of loss functions is more suitable for comparing volatility models than are the regression (3) and (4). The reason is that one forecasting model may achieve a high R^2 with values of (a, b) very different from $(0, 1)$, whereas a different model achieves a lower R^2 with parameter values, (a, b) , close to $(0, 1)$. Since the parameter estimates, \hat{a} and \hat{b} , that define the optimal affine transformation of the forecast, h_t^2 , are only known ex-post, the regressions do not provide a fair comparison.

The relevant loss function will depend on preferences, so we cannot identify a unique and natural criterion for the comparison. Rather than making a single choice, we specify six different loss functions, which can be given different interpretations. The loss functions are:

$$\text{MSE}_2 = n^{-1} \sum_{t=1}^n (\sigma_t^2 - h_t^2)^2, \quad (5)$$

$$\text{MSE}_1 = n^{-1} \sum_{t=1}^n (\sigma_t - h_t)^2, \quad (6)$$

$$\text{QLIKE} = n^{-1} \sum_{t=1}^n (\log(h_t^2) + \sigma_t^2 h_t^{-2}), \quad (7)$$

$$\text{R}^2\text{LOG} = n^{-1} \sum_{t=1}^n [\log(\sigma_t^2 h_t^{-2})]^2, \quad (8)$$

$$\text{MAD}_2 = n^{-1} \sum_{t=1}^n |\sigma_t^2 - h_t^2|, \quad (9)$$

$$\text{MAD}_1 = n^{-1} \sum_{t=1}^n |\sigma_t - h_t|. \quad (10)$$

The criteria (5), (7), and (8) were discussed by Bollerslev et al. (1994). The criteria (5) and (8) are equivalent to using the R^2 s from the regressions (3) and (4), respectively⁵, the former is also

⁴Engle & Patton (2000) also point out that heteroskedasticity of returns, r_t , implies (even more) heteroskedasticity in the squared returns, r_t^2 . So parameters are estimated inefficiently and the usual standard errors are misleading.

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

known as the mean squared forecast error criterion. The loss function (7) corresponds to the loss implied by a Gaussian likelihood, and the mean absolute deviation criteria, (9) and (10), are interesting because they are more robust to outliers than, say, the mean squared forecast error criterion.

4 Intra-Day Estimation of Volatility

Estimation of volatility models usually results in highly significant parameter estimates, as reported by numerous papers starting with the seminal paper by Engle (1982). It was therefore puzzling that volatility models could only explain a very modest amount of the out-of-sample variation of realized volatility, measured by the ex-post squared returns. This poor out-of-sample performance led several researchers to question the practical value of these models. Andersen & Bollerslev (1998a) have since refuted this skepticism by demonstrating that well-specified volatility models do provide quite accurate out-of-sample description of volatility. The previous findings were caused by the fact that r_t^2 is a noisy estimate of the volatility, and Andersen & Bollerslev (1998a) showed that the maximum obtainable R^2 from the regression (3) is indeed very small. Hence, there is not necessarily any contradiction between the highly significant parameter estimates and the poor predictive out-of-sample performance, when the squared return is used as a measure of daily volatility. Andersen & Bollerslev (1998a) suggested a more precise measure of volatility. Specifically, they show that high frequency data can be used to compute a better ex-post measure of volatility, a measure that is based on cumulative squared intra-day returns. We proceed with this idea and apply an estimate of σ_t^2 , which is based on intra-day returns.

The reason that squared returns, r_t^2 , (or squared residuals, $\hat{\epsilon}_t^2$), were substituted for σ_t^2 , was due to the fact that σ_t^2 is unobserved. There are two important reasons for using precise intra-day estimates of daily volatility when volatility models are evaluated and compared.

1. A precise estimate of σ_t^2 makes it easier to tell good models from bad models.
2. A noisy measure of σ_t^2 can severely distort the comparison and may cause an inconsistency.

The second claim follows from the following argument: Suppose that the objective is to minimize expected loss, $E [L(\sigma_t^2, h_t^2)]$. Under a sufficient set of conditions, it holds that

$$n^{-1} \sum_{t=1}^n [L(\sigma_t^2, h_{1,t}^2) - L(\sigma_t^2, h_{2,t}^2)] \xrightarrow{P} E [L(\sigma_t^2, h_{1,t}^2) - L(\sigma_t^2, h_{2,t}^2)].$$

This property allows one to use sample performances to compare models and, as n gets large, one will be able to tell which is a better model, with probability converging to one. Unfortunately, the same regularity conditions are not sufficient for

$$E [L(\sigma_t^2, h_{1,t}^2) - L(\sigma_t^2, h_{2,t}^2)] > 0$$

to imply

$$n^{-1} \sum_{t=1}^n [L(r_t^2, h_{1,t}^2) - L(r_t^2, h_{2,t}^2)] \xrightarrow{P} c > 0,$$

in general, even if r_t^2 is a well-behaved random variable that satisfies $E(r_t^2) = \sigma_t^2$. So, unless the loss function satisfies certain criteria, the substitution of r_t^2 for σ_t^2 , can cause one to conclude that an inferior model is the best forecasting model! (See Hansen & Lunde (2001b)).

We adopt a notation similar to that of Andersen & Bollerslev (1998a), in our description of intra-day data. They define the discretely observed series of continuously compounded returns with m observations per day as

$$r_{(m),t+j/m} = \log(p_{t+j/m}) - \log(p_{t+(j-1)/m}), \quad j = 1, \dots, m.$$

In this notation $r_{(1),t}$ equals the inter-day returns r_t , defined in (1), and $r_{(m),t+j/m}$ equals the return earned over a period of length $1/m$. Intra-day returns can be used to obtain a precise estimate of σ_t^2 . This can be seen from the identity

$$\begin{aligned} \sigma_t^2 &\equiv \text{var}(r_t | \mathcal{F}_{t-1}) \\ &= E \left(\sum_{j=1}^m r_{(m),t+j/m} - E(r_{(m),t+j/m} | \mathcal{F}_{t-1}) \right)^2 \\ &= \sum_{j=1}^m \text{var}(r_{(m),t+j/m} | \mathcal{F}_{t-1}) + \sum_{i \neq j} \text{cov}(r_{(m),t+i/m}, r_{(m),t+j/m} | \mathcal{F}_{t-1}). \end{aligned}$$

So provided that the intra-day returns are conditionally uncorrelated, we have the identity

$$\sigma_t^2 \equiv \text{var}(r_t | \mathcal{F}_{t-1}) = \sum_{j=1}^m \text{var}(r_{(m),t+j/m} | \mathcal{F}_{t-1}). \quad (11)$$

For large values of m , $E(r_{(m),t+j/m}|\mathcal{F}_{t-1})$ is negligible, such that

$$E(r_{(m),t+j/m}^2|\mathcal{F}_{t-1}) \simeq \text{var}(r_{(m),t+j/m}|\mathcal{F}_{t-1}). \quad (12)$$

Equations (11) and (12) motivate the use of intra-day returns to estimate σ_t^2 . If (12) holds with equality, then an unbiased estimator of σ_t^2 is given by

$$\hat{\sigma}_{(m),t}^2 \equiv \sum_{j=1}^m r_{(m),t+j/m}^2,$$

which we refer to as the m -frequency of realized daily volatility. Within the class of models we consider in this paper, $\hat{\sigma}_{(m),t}^2$ is an estimate of the conditional variance. However, in a more general setting the daily volatility is not pre-determined and $\hat{\sigma}_{(m),t}^2$ is in this case an estimate of integrated volatility, as discussed in Barndorff-Nielsen & Shephard (2001).

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 (of the m possible) returns. Assume for simplicity that we observe the f first returns, given by $r_{(m),t+j/m}^2$, $j = 1, \dots, f$. In this case we define

$$\hat{\sigma}_{(m,f),t}^2 \equiv \sum_{j=1}^f r_{(m),t+j/m}^2, \quad \text{where } f \leq m,$$

which denotes the partial m -frequency of realized volatility, (the realized volatility during the period in which we observed intra-day returns). Note that $\hat{\sigma}_{(m),t}^2 = \hat{\sigma}_{(m,m),t}^2$ and that $r_t^2 = \hat{\sigma}_{(1),t}^2 = \hat{\sigma}_{(1,1),t}^2$.

Generally, $E(\hat{\sigma}_{(m,f),t}^2) \leq E(\hat{\sigma}_{(m),t}^2)$. So $\hat{\sigma}_{(m,f),t}^2$ is not an unbiased estimator of σ_t^2 , as it only measures the volatility over a fraction, (f/m) , of the day. A simple solution would be to add the close-to-open squared returns. However this would introduce a very noisy element, similar to the inter-day squared returns, r_t^2 , and would defy the purpose of using intra-day data.

A better solution is given by Hansen & Lunde (2001b) who have shown that $\hat{\sigma}_{(m,f),t}^2$ can be scaled to provide an unbiased estimate of σ_t^2 , under fairly reasonable assumptions. The needed assumptions are:

(A.1) A constant proportion of daily conditional variance, occurs during the time where the market is open, i.e.

$$\frac{\text{var}(r_t^{open}|\mathcal{F}_{t-1})}{\text{var}(r_t^{close}|\mathcal{F}_{t-1})} = c_1 \quad (13)$$

where $r_t^{open} \equiv \sum_{j=1}^f r_{(m),t+j/m}$ is the return during the time where intra-day data are available, and $r_t^{close} \equiv r_t - r_t^{open}$ is the return for the remainder of the day.

(A.2) If $\hat{\sigma}_{(m,f),t}^2$ is a biased estimate of $\text{var}(r_t^{open}|\mathcal{F}_{t-1})$, then this bias is proportional to the conditional variance, in the sense that $E\left[\hat{\sigma}_{(m,f),t}^2\right] = c_2 \text{var}(r_t^{open}|\mathcal{F}_{t-1})$.

(A.3) The returns r_t^{open} and r_t^{close} are conditionally uncorrelated, i.e., $\text{cov}(r_t^{open}, r_t^{close}|\mathcal{F}_{t-1}) = 0$.

Hansen & Lunde (2001b) have shown that the assumptions: A.1, A.2, and A.3, imply that $E\left[\hat{\sigma}_{(m,f),t}^2\right] = c^{-1} \cdot \text{var}(r_t|\mathcal{F}_{t-1})$, where $c = (1 + c_1)/c_2$. The constant, c , can be consistently estimated by

$$\hat{c} = \left(\frac{\sum_{t=1}^n (r_t - \hat{\mu}_t)^2}{\sum_{t=1}^n \hat{\sigma}_{(m,f),t}^2} \right), \quad (14)$$

under the additional assumptions that a consistent estimator of $\mu_t = E(r_t|\mathcal{F}_{t-1})$ is available, and that a law of large numbers applies such that $n^{-1} \sum_{t=1}^n \left[(r_t - \hat{\mu}_t)^2 - \sigma_t^2 \right] \xrightarrow{p} 0$. Another result in Hansen & Lunde (2001b) is that $\text{var}(r_t^2)$ can be as much as f times $\text{var}(\hat{\sigma}_{(m,f),t}^2)$, which leaves room for a substantial improvement in using $\hat{\sigma}_{(m,f),t}^2$ as a proxy for σ_t^2 , rather than r_t^2 . This shows that by using intra-day returns to estimate the volatility, the precision of the estimation can be substantially improved.

So we apply $\hat{\sigma}_t^2 \equiv \hat{c} \cdot \hat{\sigma}_{(m,f),t}^2$ as our estimate of volatility in our comparison of models. The adjustment, \hat{c} , is only known ex-post but this should not distort our comparison, because the ex-post information is only used in the evaluation and is not included in the model's information set. If, for some reason, there is a difference between $E(\hat{\sigma}_{(m,m),t}^2|\mathcal{F}_{t-1})$ and $E(r_t^2|\mathcal{F}_{t-1})$, then the volatility models will be unable to (and are not meant to) adjust for such a bias. The volatility models are entirely based on inter-day returns, and their parameters are estimated such that they best describe the variation of (some power-transformation of) $r_t^2 = \hat{\sigma}_{(1),t}^2$. Thus, a potential difference between $E(\hat{\sigma}_{(m,m),t}^2|\mathcal{F}_{t-1})$ and $E(r_t^2|\mathcal{F}_{t-1})$ is a justification for making an adjustment, of the intra-day estimate of the volatility.

5 The Test for Superior Predictive Ability

A time-series of observations is divided into an estimation period and an evaluation period:

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

The parameters of the volatility models are estimated using the first R inter-day observations, and these parameter estimates are used to make the forecasts for the remaining n periods. During the evaluation period we estimate daily volatility using intra-day returns.

We let $l + 1$ denote the number of competing forecasting models. The k 'th model yields the sequence of forecasts

$$h_{k,1}^2, \dots, h_{k,n}^2, \quad k = 0, 1, \dots, l,$$

that are compared to the intra-day calculated volatility,

$$\hat{\sigma}_1^2, \dots, \hat{\sigma}_n^2.$$

Model k 's forecast of $\sigma_t^2, h_{k,t}^2$, leads to the observed loss $L(\hat{\sigma}_t^2, h_{k,t}^2)$, where L is defined from the performance measures listed in Section 3, e.g., $L(\hat{\sigma}_t^2, h_{k,t}^2) = (\hat{\sigma}_t^2 - h_{k,t}^2)^2$ for the mean squared forecast error criterion.

We order the models such that the first model (subscript 0) is our benchmark model. The observed performance of model k is given by $L_{k,t} \equiv L(\hat{\sigma}_t^2, h_{k,t}^2)$, and we define model k 's performance relative to that of the benchmark model as

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

In the following we assume that the vector of relative performance, $X(t) = (X_{1,t}, \dots, X_{l,t})'$, is stationary, $E|X_t|^{6+\varepsilon} < \infty$ for some $\varepsilon > 0$, and X_t is α -mixing of order $-r$ for some $r > 3(6 + \varepsilon)/\varepsilon$. These assumptions validate the use of the stationary bootstrap of Politis & Romano (1994), which we employ in our test for superior predictive ability. See Politis & Romano (1994) for details of these assumptions.

The stationary assumption allows us to define the expected relative performance of model k (relative to the benchmark),

$$\lambda_k \equiv E[X_{k,t}], \quad k = 1, \dots, l.$$

A model that outperforms the benchmark model, model k^* say, translates into a positive value of λ_{k^*} . Thus, we can analyze whether any of the competing models significantly outperform the benchmark model, by testing the null hypothesis that $\lambda_k \leq 0$, $k = 1, \dots, l$. So the null hypothesis is that none of the models are better than the benchmark. If we reject this hypothesis, we have evidence that a better model exists, relative to the benchmark model. An equivalent formulation is the following:

$$H_0 : \lambda_{\max}^s \equiv \max_{k=1, \dots, l} \frac{\lambda_k}{\sqrt{\text{var}(n^{1/2} \bar{X}_{k,t})}} \leq 0,$$

where λ_{\max}^s denotes the best standardized performance relative to the benchmark model.

We can, by the law of large numbers, estimate the parameter, λ_k , with the sample average $\bar{X}_{k,n} = n^{-1} \sum_{t=1}^n X_{k,t}$, and λ_{\max}^s is therefore consistently estimated by

$$\bar{X}_{\max,n}^s \equiv \max_{k=1, \dots, l} \frac{\bar{X}_{k,n}}{\sqrt{\widehat{\text{var}}(n^{1/2} \bar{X}_{k,n})}},$$

provided that $\widehat{\text{var}}(\sqrt{n} \bar{X}_{k,n}) \xrightarrow{p} \Omega_{kk} \equiv \text{plim}_{n \rightarrow \infty} \text{var}(\sqrt{n} \bar{X}_{k,n})$. This statistic, $\bar{X}_{\max,n}^s$ denotes the largest t -statistic of relative performance. Even if $\lambda_{\max}^s \leq 0$ it can (and will by chance) happen that $\bar{X}_{\max,n}^s > 0$. The relevant question is whether $\bar{X}_{\max,n}^s$ is too large for it to be plausible that λ_{\max}^s is truly non-positive. This is precisely what the test for SPA is designed to answer. The test for SPA estimates the distribution of $\bar{X}_{\max,n}^s$ under the null hypothesis, and applies this distribution to evaluate whether $\bar{X}_{\max,n}^s$ is too large, to be consistent with the null hypothesis. Thus, if we obtain a small p -value, we reject the null and conclude that there is a competing model, which is significantly better than the benchmark.

White (2000) developed this framework to make a test for SPA operational. His test is known as the Reality Check (RC), which applies a supremum over the non-standardized performances, $\bar{X}_{\max,n} \equiv \max_{k=1, \dots, l} \bar{X}_{k,n}$, and a conservative asymptotic distribution. Compared to the test of Hansen (2001), the RC has lower power and is sensitive to the inclusion of poor models. The lower power of the RC has two sources. One is the fact that the tails of the distribution of $\bar{X}_{\max,n}$, are defined by the models with the largest variance, the second is a distortion that poor models causes to the RC, see Hansen (2001) for more details. The improved power properties are important for our application, as increased power makes it easier to detect volatility models with superior forecasting abilities.

Compared to tests that are based on Bonferroni bounds, the SPA-test has far better power properties. The reason is that the Bonferroni bound ignores the correlation between models, which forces Bonferroni tests to be conservative, and it is therefore difficult to detect a superior model using a Bonferroni approach.

We can describe the performance of the l models, relative to the benchmark, by the l -dimensional vector $\mathbf{X}_t = (X_{1,t}, \dots, X_{l,t})'$, $t = 1, \dots, n$, such that the sample performance is given by $\bar{\mathbf{X}}_n = n^{-1} \sum_{t=1}^n \mathbf{X}_t$. The fundamental assumption that enables the test for SPA is that $\bar{X}_{\max,n}^s$ (appropriately scaled) converges in distribution. Given the assumptions stated earlier in this section, $\{\mathbf{X}_t\}$ satisfies the conditions of a central limit theorem for dependent observation, and it follows that

$$n^{1/2}(\bar{\mathbf{X}}_n - \boldsymbol{\lambda}) \xrightarrow{d} N_l(\mathbf{0}, \boldsymbol{\Omega}), \quad (15)$$

where “ \xrightarrow{d} ” denotes convergence in distribution, where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_l)'$, and where

$$\boldsymbol{\Omega} \equiv \lim_{n \rightarrow \infty} E \left[n (\bar{\mathbf{X}}_n - \boldsymbol{\lambda}) (\bar{\mathbf{X}}_n - \boldsymbol{\lambda})' \right].$$

So as $n \rightarrow \infty$, $\bar{\mathbf{X}}_n$ converges to $\boldsymbol{\lambda}$ in probability, and by Slutsky’s theorem it follows that $\bar{X}_{\max,n}^s$ converges to λ_{\max}^s in probability. Therefore, a large positive value of $\bar{X}_{\max,n}^s$ indicates that the benchmark model is outperformed. The tests for SPA (tests for superior predictive ability) of Hansen (2001) applies the result in (15) to derive a critical value for $\bar{X}_{\max,n}^s$, and this critical value is the threshold at which $\bar{X}_{\max,n}^s$ becomes too large for it to be plausible that $\lambda_{\max}^s \leq 0$.

5.1 Bootstrap Implementation

In practice it is often impossible to get a sensible estimate of the $l \times l$ covariance matrix, $\boldsymbol{\Omega}$, because the number of models, l , may be large relative to the sample size, n . It is therefore very convenient to use a bootstrap implementation, as it can circumvent the obstacle of estimating $\boldsymbol{\Omega}$. The bootstrap implementation of the test for SPA is identical to the implementation used in the Reality Check by White (2000). The principle of the bootstrap method is that it uses the observed sample to generate “random” draws from the distribution of $\bar{\mathbf{X}}_n$, where $E(\bar{\mathbf{X}}_n) = \boldsymbol{\lambda}$ satisfies the null hypothesis, i.e., $\boldsymbol{\lambda} \leq \mathbf{0}$. The number of draws is denoted by B , and these draws are used to estimate $\text{var}(\bar{X}_{k,n})$ and to approximate the distribution of $\bar{X}_{\max,n}^s$. The estimated distribution of $\bar{X}_{\max,n}^s$, is then used to derive a critical value and the p -value of the test.

We let $b = 1, \dots, B$ index the re-samples of $\{1, \dots, n\}$, given by $\theta_b(t)$, $t = 1, \dots, n$. The number of bootstrap re-samples, B , should be chosen large enough not to affect the outcome of the procedure, e.g., by applying the three-step method of Andrews & Buchinsky (2000). We apply the stationary bootstrap of Politis & Romano (1994), where $\theta_b(t)$ is constructed by combining blocks with random length, where the lengths are geometrically distributed with parameter $q \in (0, 1]$. The parameter, q , is used to preserve possible time-dependence in $X_k(t)$.⁶ The re-samples are generated as follows:

1. Initiate the random variable, $\theta_b(1)$, uniformly distributed on $\{1, \dots, n\}$.
2. For $t = 2, \dots, n$.
 Generate u uniformly on $[0, 1]$.
 - (a) If u is smaller than q , then the next observation is chosen uniformly on $\{1, \dots, n\}$, just as the initial observation was chosen.
 - (b) Otherwise, if $u \geq q$, then $\theta_b(t) = \theta_b(t-1)1_{(\theta_b(t-1) < n)} + 1$, where $1_{(\cdot)}$ is the indicator function. Thus $\theta_b(t)$ is the integer that follows the value of $\theta_b(t-1)$, except if $\theta_b(t-1) = n$, in which case $\theta_b(t) = 1$.

Thus, a re-sample generated in this way, might look like the following:

$$(\theta_b(1), \dots, \theta_b(n)) = \underbrace{(n-1, n, 1, 2, 3, 76, \dots, 47, 48)}_{n \text{ elements}}.$$

Each of the re-samples of indices defines a re-sample of the X -variables, given by

$$X_{b,k}^*(t) \equiv X_k(\theta_b(t)), \quad b = 1, \dots, B, \quad t = 1, \dots, n,$$

and the averages

$$\bar{X}_{b,k,n}^* \equiv n^{-1} \sum_{t=1}^n X_{b,k}^*(t), \quad b = 1, \dots, B.$$

⁶A small value of q correspond to long blocks, and the general consistency of the stationary bootstrap requires that $q \rightarrow 0$ as $n \rightarrow \infty$. Often it is reasonable to expect that the vector of relative forecast performances, $\mathbf{X}(t)$, is fairly close to being martingale difference processes. A moderate value of q ($q = 0.5$) is therefore sufficient to capture the autocorrelation in our analysis.

These resamples can be used to approximate the distribution of our test statistic, $\bar{X}_{\max,n}^s$, as we define

$$\bar{X}_{b,\max,n}^{*,s} \equiv \max_{k=1,\dots,l} \frac{\bar{X}_{b,k,n}^* - \hat{\lambda}_k}{\sqrt{\widehat{\text{var}}(n^{1/2} \bar{X}_{k,n})}}, \quad b = 1, \dots, B, \quad (16)$$

where $\hat{\lambda}_k$ is an estimate of λ_k , which is consistent with the null hypothesis, i.e., $\hat{\lambda}_k \leq 0$, and where

$$\widehat{\text{var}}(n^{1/2} \bar{X}_{k,n}) \equiv \frac{n}{B} \sum_{b=1}^B (\bar{X}_{b,k,n}^* - \bar{X}_{k,n})^2.$$

Under our assumptions, Politis & Romano (1994) have shown that the empirical distribution of the bootstrap resamples, $\bar{\mathbf{X}}_{b,n}^* = (\bar{X}_{b,1,n}^*, \dots, \bar{X}_{b,l,n}^*)'$, converges to the true asymptotic distribution of $\bar{\mathbf{X}}_n = (\bar{X}_{1,n}, \dots, \bar{X}_{l,n})'$. In our test for SPA, we impose the empirical distribution to satisfy the null hypothesis. This enables us to test the null hypothesis. If indeed the null hypothesis is true, then $\bar{\mathbf{X}}_n$ is unlikely to be an extreme observation in the empirical distribution of the bootstrap variables, $\bar{\mathbf{X}}_{1,n}, \dots, \bar{\mathbf{X}}_{B,n}$, and consequently $\bar{X}_{\max,n}^s$ is unlikely to be extreme relative to the $\bar{X}_{1,\max,n}^{*,s}, \dots, \bar{X}_{B,\max,n}^{*,s}$.

It is the estimator of λ_k in equation (16) that ensures that the empirical distribution of the bootstrap variables, $\bar{X}_{b,\max,n}^{*,s}$, is consistent with the null hypothesis. The null hypothesis is a composite hypothesis (many values of λ are consistent with H_0) and the choice of $\hat{\lambda}_k$ is by no means unique. A conservative choice is to use $\hat{\lambda}_k = 0$, which leads to the SPA_u test whereas a liberal choice is $\hat{\lambda}_k = \min(\bar{X}_{k,n}, 0)$, which leads to the SPA_l test. The subscripts refer to the upper and lower, as the SPA_l and SPA_u provide lower and upper bounds for the p -value. The SPA_c applies

$$\hat{\lambda}_k = \begin{cases} \bar{X}_{k,n} & \text{if } \bar{X}_{k,n} \leq -A_{k,n} \\ 0 & \text{otherwise,} \end{cases}$$

where $A_{k,n}$ is given below, and this results in a consistent estimate of the p -value. The reason, for the correction fraction is that we need to determine which models have $\lambda_k = 0$, in a (strongly) consistent way. The asymptotic distribution of the test statistic depends on the value of λ , in a way that only the models with $\lambda_k = 0$ matters, and the correction above ensures that we obtain consistent p -values, see Hansen (2001) for details.

The SPA_u's choice of $\hat{\lambda}_k$ results in a conservative test and its p -value can be viewed as an upper bound for the true p -value. The consistent p -value of the SPA_c is achieved by the correction factor, $A_{k,n}$, which must be constructed such that it vanishes asymptotically, $A_{k,n} \xrightarrow{p} 0$. However, the rate at which it vanishes must be slow enough such that, as $n \rightarrow \infty$, we are able to determine exactly the models that have $\lambda_k = 0$. This is important for the consistency, because the models with $\lambda_k < 0$ do not influence the distribution of $\bar{X}_{\max,n}$ in the limit, see Hansen (2001). Even though both the SPA_l and the SPA_c apply consistent estimators for λ_k under the null hypothesis, only the SPA_c achieves generally consistent p -values. The SPA_u applies the same asymptotic distribution as the Reality Check of White (2000), which results in inconsistent p -values, unless $\lambda = \mathbf{0}$, that is, if all models have the same expected performance.

As previously noted, the correction factor, $A_{k,n}$, needs to converge to zero almost surely, at a sufficiently slow rate. The correction suggested in Hansen (2001) is given by

$$A_{k,n} \equiv \frac{1}{4} n^{-1/4} \sqrt{\widehat{\text{var}}(n^{1/2} \bar{X}_{k,n})}. \quad (17)$$

Simpler choices are available, for example $A_{k,n} = n^{-1/4}$ is an alternative choice. But it is convenient to let the correction factor depend on the individual models, because it can result in better small sample properties.

From the bootstrap resamples, we generated draws of $\bar{X}_{n,\max}^s$, given by $\bar{X}_{1,\max,n}^{*,s}, \dots, \bar{X}_{B,\max,n}^{*,s}$. We can evaluate whether $\bar{X}_{n,\max}^s$ is an extreme observation or not, by calculating the SPA p -value, which is given by

$$\hat{p}_{\text{SPA}} \equiv \sum_{b=1}^B \frac{1(\bar{X}_{b,\max,n}^{*,s} > \bar{X}_{\max,n}^s)}{B},$$

where $1(\cdot)$ is the indicator function. If $\bar{X}_{\max,n}^s$ is an extreme observation, then relatively few (possibly none) of the bootstrap draws $\bar{X}_{b,\max,n}^{*,s}$ are larger than $\bar{X}_{\max,n}^s$, and this results in a low p -value. A low p -value translates into evidence against the null hypothesis, and hence, evidence that a better forecasting model is available.

This procedure is done with each of the three estimates of the distribution, by which we obtain a lower and an upper bound for the p -value, as well as a consistent estimate of the p -value. Small sample properties of p -values obtained with the SPA_c test will depend on the actual choice of correction factors $A_{k,n}$, $k = 1, \dots, l$. It is therefore convenient to accompany a

consistent p -value with an upper and lower bound, unless the sample size is large. In a situation where n is large, or where both the upper and lower bound of the p -value point to the same conclusion, one need not worry about lack of uniqueness of the correction factor, $A_{k,n}$.

6 Empirical Results

The models are estimated using inter-day returns over the sample period $t = -R + 1, \dots, 0$, and intra-day returns are used to construct a precise estimate of the conditional variance, for each day in the evaluation period, $t = 1, \dots, n$.

6.1 Estimation of Volatility Models

All models were estimated using the method of maximum likelihood. The optimization algorithm was implemented using C++, in which the likelihood functions were maximized using the simplex method described in Press, Teukolsky, Vetterling & Flannary (1992). A total of 330 models were estimated⁷.

Because the likelihood function is rather complex for most of the volatility models, it can be difficult for general maximization routines to determine the global optimum. However, in this situation where we estimate a large number of models, some of which are quite similar, we can often provide the maximization routine with good starting values of the parameters, to ease the estimation. Nevertheless, given the large number of models and the complex structure of their likelihood functions, it is possible that the algorithm failed to maximize the likelihood function for one or more models. But we did not notice any obvious inconsistencies and for nested models we have checked that the maximum value of the likelihood function is larger for the more general model.

These models were estimated to fit two data sets. The first data set consists of daily returns for the DM-\$ spot exchange rate from October 1, 1987, through September 30, 1992 – a total of 1,254 observations. This data set has previously been analyzed by Andersen & Bollerslev

⁷Due to space constraints we have not included all of our results. An extensive collection of our results are given in a technical appendix, which interested readers are referred to. The appendix can be downloaded from <http://www.hha.dk/~alunde/academic/research/papers/vola-mod-appendix.pdf>.

(1998a). The second data set contains daily returns from closing prices on the IBM stock from January 2, 1990, through May 28, 1999 – a total of 2,378 observations.

6.2 Realized Volatilities for the Exchange Rate Data

Our out-of-sample exchange rate data⁸ are identical to the data used in Andersen & Bollerslev (1998a). Our estimation of realized volatility is based on temporal aggregates of five-minute returns; this corresponds to $m = 288$. The out-of-sample DM-\$ exchange rate data covers the period from October 1, 1992, through September 30, 1993. This results in a total of 74,880 five-minute returns, and volatility estimates for 260 days. Using $r_{(288),t}$, our 288-frequency estimate of volatility is denoted by $\hat{\sigma}_{(288),t}^2$. It is the (adjusted) ex-post measure of volatility that is compared to the models' forecast of volatility, denoted by $h_{k,t}^2$, $k = 1, \dots, l$. The significance of relative performance across models is then evaluated using the test for SPA.

In the technical appendix, Hansen & Lunde (2001a), we list the R^2 s (denoted R_{inter}^2 and R_{intra}^2) from the regressions corresponding to (3) and (4) for $m = 1$, and $m = 288$, that is

$$\hat{\sigma}_{(1),t}^2 = a + bh_{k,t}^2 + u_t \quad (18)$$

$$\hat{\sigma}_{(288),t}^2 = a + bh_{k,t}^2 + u_t. \quad (19)$$

We find that R_{inter}^2 is typically between 2 and 4 per cent, a very small figure compared to R_{intra}^2 , which typically lies between 35 and 45 per cent. We also computed the R^2 (denoted R_{inter}^{*2} and R_{intra}^{*2}) from the log-regression (4). This generally resulted in smaller values of the R^2 s, but the large difference between the intra-day and the inter-day measure persisted. The adjusted estimates of volatilities are given by $\hat{\sigma}_t^2 = \hat{c} \cdot \hat{\sigma}_{(288),t}^2$, where $\hat{c} = .8418$ is the estimate defined in (14). The returns and the intra-day estimates of volatility are plotted in Figure 2.

6.3 Realized Volatilities for the IBM Data

These data were extracted from the Trade and Quote (TAQ) database. The TAQ database is a collection of all trades and quotes in the New York Stock Exchange (NYSE), American Stock

⁸We thank Tim Bollerslev providing us with the intra-day exchange rate data. For the construction of the series and additional information, we refer to Andersen & Bollerslev (1997, 1998b) and Andersen, Bollerslev, Diebold & Labys (2000)

Exchange (AMEX), and National Association of Securities Dealers Automated Quotation (Nasdaq) securities. In our estimation of intra-day volatility, we only included trades and quotes from the NYSE. Schwartz (1993) and Hasbrouck, Sofianos & Sosebee (1993) document NYSE trading and quoting procedures. This out-of-sample series runs from June 1, 1999, through May 31, 2000, spanning a total of 254 trading days.

As noted by several authors, it is important to take market micro-structures into account. Factors, such as the bid-ask bounds and the irregular spacing of price quotes, could potentially distort our estimates of volatility, for example unadjusted estimates based on tick-by-tick data are likely to be biased. Andersen & Bollerslev (1997, 1998a, 1998b) and Andersen, Bollerslev, Diebold & Ebens (2000) circumvent this obstacle by estimating the volatility from artificially constructed five-minute returns⁹. We take a similar approach, in the sense that we fit a cubic spline through all mid-quotes of a given trading day from the time interval 9:30 EST – 16:00 EST. This is done by applying the **Splus** routine called **smooth-spline**¹⁰. A random sample of these splines, as well as mid quotes, are displayed in Figure 1. From the splines we extract artificial three-minute returns, which leads to $f = 130$ three-minute returns for each of the days. This delivers our measure of realized volatility. Because we only have 130 of the 480 theoretical three-minute returns, we denote our measure for the volatility by

$$\hat{\sigma}_{(m,f),t}^2 = \sum_{j=1}^f r_{(m),t+j/m}^2,$$

where $(m, f) = (480, 130)$ for the three-minute returns.

We computed the R^2 s for this data set. The relationship between R_{inter}^2 and R_{intra}^2 , and R_{inter}^{*2} and R_{intra}^{*2} were analogous to the exchange series but the R^2 s were somewhat lower. R_{intra}^2 ranged between 2 and 15 per cent, again in contrast to R_{inter}^2 , which in all cases was below 1.25 per cent.

The intra-day measure $\hat{\sigma}_{(480,130),t}^2$ is not directly comparable to the inter-day measure, $\hat{\sigma}_{(1),t}^2$, because they are calculated from the part of the day where intra-day data are available. So,

⁹Other ways of estimating the daily volatility have been suggested. The linear interpolation and previous-tick methods are described in Dacorogna, Gencay, Müller, Olsen & Pictet (2001), and a Fourier based method has been suggested by Malliavin & Mancino (2002) and applied to high-frequency data by Barucci & Reno (2001). In the next section we show that our result are robust with respect to the method of constructing intra-day returns.

¹⁰This is a one-dimensional cubic smoothing spline which uses a basis of B-splines as discussed in chapters 1,2 & 3 of Green & Silverman (1994).

we need to adjust for this bias in order to avoid a distortion of the evaluation based on the loss functions (5–10).

It is interesting to note that this bias does not affect the R^2 s obtained from (3) and (4), because the R^2 is invariant to affine transformations $x \mapsto a+bx$, provided that $b \neq 0$. However, this reveals a shortcoming of using the R^2 for evaluation of forecasting models. A model that consistently has predicted the volatility to be half of what the realized volatility turned out to be, would obtain a perfect R^2 of 1, whereas a model that on average is better at predicting the level of the volatility, yet not perfectly, would obtain an R^2 less than one. If one were to make a strict comparison of the two models, then clearly the latter is a better choice, and the R^2 is misinformative in this case. Thus, if the R^2 is better for one model compared to another, it only tells us that there is an affine transformation of the “better” model, which is better than any affine transformation of the other model. Since the “optimal” affine transformation is only known ex-post, it is not necessarily a good criterion for comparison of volatility models.

The volatility estimates based on the three-minute returns need to be adjusted by $\hat{c} = 4.4938$, which is a more than $\frac{480}{130} \simeq 3.7$, which corresponds to the length of a day, relative to the fraction of the day we have intra-day observations. Thus, the squared three-minute returns (from the proportion of the day we have intra-day returns) underestimated the daily volatility, by a factor of about $4.5/3.7$. The estimated intra-day volatilities are plotted in Figure 3 along with the daily returns.

There are several possible explanations, for 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 autocorrelation in the intra-day returns can cause a bias. This can be seen from the relation

$$r_t^2 = \sum_{j=1}^m r_{t+j/m}^2 + \sum_{i \neq j} r_{t+i/m} r_{t+j/m}.$$

If we ignore that only a fraction of the intra-day returns are observed, we have evidence that $\sum_{t=1}^n r_t^2 > \sum_{t=1}^n \left(\sum_{j=1}^m r_{t+j/m}^2 \right)$, which implies that the last term $\sum_{t=1}^n \left(\sum_{i \neq j} r_{t+i/m} r_{t+j/m} \right)$ is positive. A “positive average correlation” can arise from the market micro-structures, but it may also be an artifact of the way we construct the artificial intra-day returns. These are created

by fitting a number of cubic splines to the data, and if this spline method results in an over-smoothing of the intra-day data, it will result in a positive correlation. 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 is in conflict with the findings of French & Roll (1986) and Baillie & Bollerslev (1989), so we find this explanation to be unrealistic. Finally, a fourth factor that can create a difference between squared inter-day returns and the sum of squared intra-day returns, is the neglect of the conditional expected value $E(r_{t+i/m}|\mathcal{F}_{t-1})$, $i = 1, \dots, m$. Suppose that $E(r_{t+i/m}|\mathcal{F}_{t-1}) = 0$ for $i = 1, \dots, f$, 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 intra-day 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, because adjustments cannot be made to a portfolio while the market is closed.

As described above, it is not important which of the four explanations cause the difference, as long as our adjustment does not favor some models over others. It is unlikely that the adjustment should favor some models over others, as the adjustment is made ex-post, and is made independent of the model forecasts. So the adjustment should not matter for our comparison. To verify that our results are not influenced by the cubic spline method we have use to construct the intra-day returns, we have repeated the analysis using several other methods to construct intra-day returns, and found our results to be robust.

6.4 Results from the Model Comparison

The models were compared using two different benchmark models, a GARCH(1, 1) model and an ARCH(1) model. The ARCH(1) model serves as a pseudo benchmark model, which is included as a point of reference and to verify that the test for SPA has power. The ARCH(1) is expected to be a poor model, so the SPA test should be able to detect that it is an inferior model.

Our results are given in Tables 3 and 4.¹¹ To verify that our results are robust to the choice

¹¹The results are based on $B = 2,000$ bootstrap replications using $q = 0.5$ as the bootstrap dependence-parameter.

of intra-day estimation of volatilities, the same analysis was made using six other estimators. These results are presented in Table 6. It is clear that the ARCH(1) model is outperformed by alternative models, for all loss functions and in both data sets. The GARCH(1, 1) model does quite well in the exchange rate data, and the GARCH model is not significantly outperformed by other models in this sample. Notice how much the SPA p -values are increased when the GARCH(1, 1) is used as the benchmark model, instead of the ARCH(1). This is due to the better performance by the GARCH(1, 1) compared to the ARCH(1) model. In the analysis of the IBM data, there is strong evidence that the GARCH(1, 1) is inferior to other models.

Table 5 contains the p -values, had we used a less powerful test statistic, which take the supremum over relative performances without scaling the performance by its standard deviation. The p -values reported in the right most column, are the p -values that the Reality Check of White (2000) would produce. As can be seen, there is less evidence against the GARCH(1, 1) had we used this test statistic, this is due to inferior power properties this test statistic results in. It is also interesting to note that the p -values of the three tests for SPA differ in some cases. The difference between the three tests is caused by the choice of asymptotic distribution, which the tests apply under the null hypothesis. The SPA_l and SPA_u both provide inconsistent p -values, however they are useful as they provide bounds for the consistent p -value (that of the SPA_c). The SPA_u is sensitive to inclusion of models that are worse than the benchmark, and this property is made clear in Table 3, where the three p -values agree when the ARCH is used as the benchmark, but differs substantially when the GARCH(1, 1) is used as the benchmark, see Hansen (2001) for more details.

The test for superior predictive ability can evaluate whether better forecasting models exist, but does not identify the best model, as it is not a model selection criterion. Additional information about the relative ranking of the models is listed in Tables 7 and 8, which contain the result for the models with a zero-mean specification. Results for the two other mean specifications are given in the technical appendix. The scores in these tables denote the percentage of models (out of the 330 models) that performed worse than a given model (given from the row), using a particular loss function and a particular data set (given from the column). Thus the best, worst, and median performing models have the scores 100, 0, and 50 respectively. Since we have six criteria and two data sets, each model has 12 scores. The last column in the tables is the average

of the 12 scores.

As can be seen from the Tables 7 and 8, the ARCH(1) model is generally one of the worst models. However, in the analysis of the IBM data, there are about 25% of the volatility models that perform worse than the ARCH(1), if the mean absolute criterion is applied. It is interesting that this high a percentage of the far more sophisticated models are performing worse than the simple ARCH(1) model in this respect. The GARCH(1, 1) model does quite well in the exchange rate data, but falls behind in the IBM data. It is interesting to notice that it is not the same models that do well in the two data sets, nor do the different criteria point to the same models as the better models.

In the exchange rate data set, the best models are GARCH(2, 2), the LOG-GARCH(2, 2), and the GQ-ARCH(2,1) models. In terms of combinations of error distribution and mean function there is not a clear winner, and we have only listed the model with a constant mean specification in Tables 7 and 8. The overall best GARCH(2, 2) model is the one with t -distributed errors and GARCH-in-mean, see the technical appendix, the overall best LOG-GARCH(2, 2) model is the model with Gaussian errors and a zero-mean specification, see Tables 7. The best GQ-ARCH(2,1) model is the model with Gaussian errors and GARCH-in-mean. See also Figures 4–7, which show the population of model performances. Each figure corresponds to a particular loss function (and data set) and contains four panels. The density of the performance of the 330 forecasting models is given in the upper left panel. The last three panels show the performance densities after grouping the models according to the specifications: Gaussian versus t -distributed specification, models with and without a leverage effect, and the three mean specification. These three groupings of the models are shown in the upper-right, lower-left, and lower-right panel respectively.¹²

The Gaussian specification typically does better than the t -distributed specification of standardized returns. However, for the less outlier sensitive loss functions (MAD_1 and MAD_2) we see that the very best model is always a model with a t -distribution specification. Interestingly, in the analysis of the the exchange rate data, the models that can accommodate a leverage effect do worse than those without, whereas the opposite is the case in the analysis the IBM data. The

¹²To save space, we have only included the figures that correspond to the MSE_2 and MAD_2 loss functions. All figures can be found in Hansen & Lunde (2001a).

mean specifications are almost identical for all loss functions and for both data sets. Although the conditional mean $\mu_t = E(r_t|\mathcal{F}_{t-1})$ is likely to be small, this result was not obvious a priori. A small improvement of the modelling of the conditional mean, may lead to clear improvements in the forecast of volatility. On the other hand, additional parameters introduced in the mean-specification, could have led to worse forecast overall.

When analyzing the IBM data it is more clear which is a better model. The best overall performing model is the A-PARCH(2, 2) model of Ding et al. (1993), specified with t -distributed errors and mean zero, see Table 8. Also the V-GARCH with a Gaussian, GARCH-in-mean specification does quite well in terms of the two MAD criteria, which are less sensitive to outliers. It is also interesting that all the EGARCH(p, q) models with Gaussian errors are relatively poor, except for the model that has $(p, q) = (1, 2)$. Notice how much lower the model with $(p, q) = (2, 2)$ is ranked. A plausible explanation for this drop in the ranking, as an extra lag is added to the model, is that the more general model overfits the in-sample observation, which hurts the model in the out-of-sample evaluation. The fact that the EGARCH specification performs far better using t -distributed standardized errors compared to Gaussian, shows the importance of modelling the entire distribution. It is not sufficient to focus on the specification of the volatility, even if the volatility is the only object of interest. The IGARCH models are surprisingly poor, and it shows that the restrictions these models impose are likely to be invalid.

As can be seen from Figure 6 and Figure 7 in particular, the models with the best sample performance, in the analysis of IBM data, are models that can accommodate a leverage effect.

7 Summary and Concluding Remarks

We have compared a large number of volatility models, which are estimated using inter-day returns. The estimated models are compared in terms of their out-of-sample ability to describe the variation in volatility. This out-of-sample comparison leads to a one-step-ahead comparison of the models forecast of conditional variance, which we compare to intra-day estimated measures of volatility. The intra-day estimated volatilities provide good estimates of the daily volatility, which makes the comparison of different volatility models more powerful.

The performances of the volatility models were measured using a number of different loss

functions, and the significance of the different performances of the models was evaluated using the test for superior predictive ability of Hansen (2001). Our empirical analysis showed the usefulness of this test, and illustrated the superior power properties it has compared to related tests. The test for superior predictive ability is not a model selection criterion, and we did not attempt to identify the best volatility model. However the use of significance tests, rather than a model selection criterion, allows us to make strong conclusion about which models are inferior.

Our analysis was limited to DM-\$ exchange rates, and IBM stock returns and the use of 330 different forecasting models, yet we obtained several interesting results. There is no evidence that the GARCH(1, 1) model is outperformed in our analysis of the exchange rate data. This cannot be explained by the SPA test being entirely powerless, as the ARCH(1) model was clearly rejected. In the analysis of IBM stock returns we found conclusive evidence that the GARCH(1, 1) is an inferior model, and it was mostly models that can accommodate a leverage effect that had a better sample performance.

The fact that volatility models are capable at predicting intra-day volatility fairly accurately, is an accomplishment in itself, since these volatility models are based on a relatively small information set (containing inter-day returns). Nevertheless, it is questionable whether the modest gains that we found more complicated volatility models have, relative to the GARCH(1, 1) model, are sufficiently large to justify the resources that researchers have devoted to the constructions of the many GARCH-type models.

Some new directions in the construction of volatility models seem more promising to us. For example, models that incorporate intra-day returns in their information set, such as the recent VAR approach of Andersen et al. (2001) may outperform any model, that is based on inter-day returns. The class of the stochastic volatility models may also lead to a better description of daily volatility. In this paper we used intra-day returns to estimate daily volatility. But it is likely that intra-day returns, which are now readily available, are going to prove themselves useful for many other purposes. They are likely to be useful to derive forecasts of the entire distribution of r_t , (not just than σ_t^2), which is important for Value-at-Risk models and many financial applications. A comparison between models based on inter-day data versus intra-day based is an interesting area for future research.

References

- Andersen, T., Bollerslev, T., Diebold, F. X. & Labys, P. (2001), 'Modeling and forecasting realized volatility', *NBER Working Paper 8160*.
- Andersen, T. G. & Bollerslev, T. (1997), 'Intraday periodicity and volatility persistence in financial markets', *Journal of Empirical Finance* **4**, 115–158.
- Andersen, T. G. & Bollerslev, T. (1998a), 'Answering the skeptics: Yes, standard volatility models do provide accurate forecasts', *International Economic Review* **39**(4), 885–905.
- Andersen, T. G. & Bollerslev, T. (1998b), 'Deutsche mark-dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies', *Journal of Finance* **53**(1), 219–265.
- Andersen, T. G., Bollerslev, T., Diebold, F. X. & Ebens, H. (2000), 'The distribution of stock return volatility', *Forthcoming Journal of Financial Economics*.
- Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2000), 'The distribution of exchange rate volatility', *Journal of the American Statistical Association* **96**(453), 42–55.
- Andrews, D. W. K. & Buchinsky, M. (2000), 'A three-step method for choosing the number of bootstrap repetitions', *Econometrica* **68**, 23–52.
- Baillie, R. T. & Bollerslev, T. (1989), 'The message in daily exchange rates: A conditional variance', *Journal of Business & Economic Statistics* **7**(4), 297–305.
- Barndorff-Nielsen, O. E. & 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 B* **63**(2), 167–241.
- Barucci, E. & Reno, R. (2001), 'On measuring volatility and GARCH forecasting performance'. Working paper, Dipartimento di Economia Politica, Università di Siena.
- 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. (1987), 'Modelling persistence in conditional variances', *Econometric Reviews* **5**, 1–50.
- Bollerslev, T., Engle, R. F. & Nelson, D. (1994), ARCH models, in R. F. Engle & D. L. McFadden, eds, 'Handbook of Econometrics', Vol. IV, Elsevier Science B.V., pp. 2961–3038.
- Dacorogna, M. M., Gencay, R., Müller, U., Olsen, R. B. & Pictet, O. V. (2001), *An Introduction to High-Frequency Finance*, Academic Press.

- Diebold, F. X. & Lopez, J. A. (1996), Forecast evaluation and combination, in G. S. Maddala & C. R. Rao, eds, 'Handbook of Statistics', Vol. 14: Statistical Methods in Finance, North-Holland, Amsterdam, pp. 241–268.
- Diebold, F. X. & Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business and Economic Statistics* **13**, 253–263.
- Ding, Z., Granger, C. W. J. & Engle, R. F. (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, R. F. (1982), 'Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation', *Econometrica* **45**, 987–1007.
- Engle, R. F., Lilien, D. V. & Robins, R. P. (1987), 'Estimating time varying risk premia in the term structure: The ARCH-M model', *Econometrica* **55**, 391–407.
- Engle, R. F. & Ng, V. (1993), 'Measuring and testing the impact of news on volatility', *Journal of Finance* **48**, 1747–1778.
- Engle, R. F. & Patton, A. J. (2000), What good is a volatility model? Manuscript at Stern, NYU, <http://www.stern.nyu.edu/~rengle/papers/vol_paper_29oct.001.pdf>.
- Figlewski, S. (1997), 'Forecasting volatility', *Financial Markets, Institutions & Instruments* **6**(1), 1–88.
- French, K. R. & Roll, R. (1986), 'Stock return variance: The arrival of information and the reaction of traders', *Journal of Financial Economics* **17**, 5–26.
- Gallant, A. R. & Tauchen, G. (1989), 'Seminonparametric estimation of conditionally constrained heterogeneous processes: Asset pricing applications', *Econometrica* **57**, 1091–1120.
- Geweke, J. (1986), 'Modelling persistence in conditional variances: A comment', *Econometric Reviews* **5**, 57–61.
- Glosten, L. R., Jagannathan, R. & Runkle, D. E. (1993), 'On the relation between the expected value and the volatility of the nominal excess return on stocks', *Journal of Finance* **48**, 1779–1801.
- Green, P. J. & Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall.
- Hansen, B. E. (1994), 'Autoregressive conditional density models', *International Economic Review* **35**(3), 705–730.
- Hansen, P. R. (2001), 'An unbiased and powerful test for superior predictive ability'. Brown University, Department of Economics Working Paper, 01-06.
<http://chico.pstc.brown.edu/~phansen>.

- Hansen, P. R. & Lunde, A. (2001a), 'Technical appendix: A forecast comparison of volatility models: Does anything beat a GARCH(1,1)'.
<http://www.hha.dk/~alunde/academic/research/papers/vola-mod-appendix.pdf>.
- Hansen, P. R. & Lunde, A. (2001b), 'Volatility estimation using high frequency data with partial availability', *Mimeo*.
- Harvey, C. R. & Siddique, A. (1999), 'Autoregressive conditional skewness', *Journal of Financial and Quantitative Analysis* **34**(4), 465–487.
- Hasbrouck, J., Sofianos, G. & Sosebee, D. (1993), Orders, trades, reports and quotes at the new york stock exchange, Technical report, NYSE, Research and Planning Section.
- Hentschel, L. (1995), 'All in the family: Nesting symmetric and asymmetric garch models', *Journal of Financial Economics* **39**, 71–104.
- Higgins, M. L. & Bera, A. K. (1992), 'A class of nonlinear ARCH models', *International Economic Review* **33**, 137–158.
- Lopez, J. A. (2001), 'Evaluation of predictive accuracy of volatility models', *Journal of Forecasting* **20**(1), 87–109.
- Malliavin, P. & Mancino, M. E. (2002), 'Fourier series method for measurement of multivariate volatilities', *Forthcoming in Finance and Stochastics* **5**(1).
- Nelson, D. B. (1991), 'Conditional heteroskedasticity in asset returns: A new approach', *Econometrica* **59**, 347–370.
- Pagan, A. R. & Schwert, G. W. (1990), 'Alternative models for conditional volatility', *Journal of Econometrics* **45**, 267–290.
- Pantula, S. G. (1986), 'Modelling persistence in conditional variances: A comment', *Econometric Reviews* **5**, 71–74.
- Politis, D. N. & Romano, J. P. (1994), 'The stationary bootstrap', *Journal of the American Statistical Association* **89**, 1303–1313.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannary, B. P. (1992), *Numerical Recipes in C*, 2 edn, Cambridge University Press.
- Schwartz, R. A. (1993), *Reshaping the Equity Markets*, Business One Irwin.
- Schwert, G. W. (1989), 'Why does stock volatility change over time?', *Journal of Finance* **44**(5), 1115–1153.
- Schwert, G. W. (1990), 'Stock volatility and the crash of '87', *Review of Financial Studies* **3**(1), 77–102.
- Sentana, E. (1995), 'Quadratic ARCH models', *Review of Economic Studies* **62**(4), 639–661.

- Tauchen, G. (2001), 'Notes on financial econometrics', *Journal of Econometrics* **100**, 57–64.
- Taylor, S. J. (1986), *Modelling Financial Time Series*, John Wiley & Sons.
- West, K. D. (1996), 'Asymptotic inference about predictive ability', *Econometrica* **64**, 1067–1084.
- White, H. (2000), 'A reality check for data snooping', *Econometrica* **68**, 1097–1126.
- Zakoian, J.-M. (1994), 'Threshold heteroskedastic models', *Journal of Economic Dynamics and Control* **18**, 931–955.

Table 1: Alternative GARCH-type models: The conditional mean.

Zero mean:	$\mu_t = 0$
Non-zero constant mean:	$\mu_t = \mu_0$
GARCH-in-mean (σ^2)	$\mu_t = \mu_0 + \mu_1 \sigma_{t-1}^2$

Table 2: Alternative GARCH-type models: The conditional variance

ARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i \varepsilon_{t-i}^2$
GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$
IGARCH	$\sigma_t^2 = \omega + \varepsilon_{t-1}^2 + \sum_{i=2}^p \alpha_i (\varepsilon_{t-i}^2 - \varepsilon_{t-1}^2) + \sum_{j=1}^q \beta_j (\sigma_{t-j}^2 - \varepsilon_{t-1}^2)$
Taylor/Schwert:	$\sigma_t = \omega + \sum_{i=1}^p \alpha_i \varepsilon_{t-i} + \sum_{j=1}^q \beta_j \sigma_{t-j}$
A-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^p [\alpha_i \varepsilon_{t-i}^2 + \gamma_i \varepsilon_{t-i}] + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$
NA-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i (\varepsilon_{t-i} + \gamma_i \sigma_{t-i})^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$
V-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i (e_{t-i} + \gamma_i)^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$
Thr.-GARCH:	$\sigma_t = \omega + \sum_{i=1}^p \alpha_i [(1 - \gamma_i) \varepsilon_{t-i}^+ - (1 + \gamma_i) \varepsilon_{t-i}^-] + \sum_{j=1}^q \beta_j \sigma_{t-j}$
GJR-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{p_1} [\alpha_i + \gamma_i I_{\{\varepsilon_{t-i} > 0\}}] \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$
log-GARCH:	$\log(\sigma_t) = \omega + \sum_{i=1}^p \alpha_i e_{t-i} + \sum_{j=1}^q \beta_j \log(\sigma_{t-j})$
EGARCH:	$\log(\sigma_t^2) = \omega + \sum_{i=1}^p [\alpha_i e_{t-i} + \gamma_i (e_{t-i} - E e_{t-i})] + \sum_{j=1}^q \beta_j \log(\sigma_{t-j}^2),$
NGARCH ^a :	$\sigma_t^\delta = \omega + \sum_{i=1}^p \alpha_i \varepsilon_{t-i} ^\delta + \sum_{j=1}^q \beta_j \sigma_{t-j}^\delta$
A-PARCH:	$\sigma_t^\delta = \omega + \sum_{i=1}^p \alpha_i [\varepsilon_{t-i} - \gamma_i \varepsilon_{t-i}]^\delta + \sum_{j=1}^q \beta_j \sigma_{t-j}^\delta$
GQ-ARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i \varepsilon_{t-i} + \sum_{i=1}^p \alpha_{ii} \varepsilon_{t-i}^2 + \sum_{i < j}^p \alpha_{ij} \varepsilon_{t-i} \varepsilon_{t-j} + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$
H-GARCH:	$\sigma_t^\delta = \omega + \sum_{i=1}^p \alpha_i \delta \sigma_{t-i}^\delta [e_t - \kappa - \tau (e_t - \kappa)]^\nu + \sum_{j=1}^q \beta_j \sigma_{t-j}^\delta$
Aug-GARCH ^b :	$\sigma_t^2 = \begin{cases} \delta \phi_t - \delta + 1 ^{1/\delta} & \text{if } \delta \neq 0 \\ \exp(\phi_t - 1) & \text{if } \delta = 0 \end{cases}$ $\phi_t = \omega + \sum_{i=1}^p [\alpha_{1i} \varepsilon_{t-i} - \kappa ^\nu + \alpha_{2i} \max(0, \kappa - \varepsilon_{t-i})^\nu] \phi_{t-j}$ $+ \sum_{i=1}^p [\alpha_{3i} f(\varepsilon_{t-i} - \kappa , \nu) + \alpha_{4i} f(\max(0, \kappa - \varepsilon_{t-i}), \nu)] \phi_{t-j}$ $+ \sum_{j=1}^q \beta_j \phi_{t-j}^2$

^a This is A-PARCH without the leverage effect.

^b Here $f(x, \nu) = (x^\nu - 1)/\nu$.

Table 3: Exchange Rate Data (DM/USD)

Criterion	Benchmark: ARCH(1)							
	Performance				p -values			
	Bench.	Worst	Median	Best	Naive	SPA _l	SPA _c	SPA _u
MSE ₂	-0.1288	-0.1404	-0.0853	-0.0778	0.0420	0.0845	0.0910	0.0950
MSE ₁	-0.0463	-0.0492	-0.0339	-0.0314	0.0085	0.0180	0.0180	0.0225
QLIKE	-0.3747	-0.3765	-0.3332	-0.3252	0.0080	0.0175	0.0190	0.0195
R ² LOG	-0.4124	-0.4250	-0.3366	-0.3154	0.0005	0.0005	0.0005	0.0005
MAD ₂	-0.2533	-0.2904	-0.2194	-0.2045	0.0010	0.0030	0.0035	0.0040
MAD ₁	-0.1698	-0.1834	-0.1473	-0.1396	0.0000	0.0005	0.0005	0.0005

Criterion	Benchmark: GARCH(1, 1)							
	Performance				p -values			
	Bench.	Worst	Median	Best	Naive	SPA _l	SPA _c	SPA _u
MSE ₂	-0.0812	-0.1404	-0.0853	-0.0778	0.1975	0.6045	0.7760	0.9370
MSE ₁	-0.0321	-0.0492	-0.0339	-0.0314	0.2870	0.3235	0.4695	0.7945
QLIKE	-0.3280	-0.3765	-0.3332	-0.3252	0.2655	0.5835	0.7755	0.9605
R ² LOG	-0.3218	-0.4250	-0.3366	-0.3154	0.0760	0.2280	0.3280	0.6625
MAD ₂	-0.2107	-0.2904	-0.2194	-0.2045	0.1695	0.2115	0.2900	0.5760
MAD ₁	-0.1415	-0.1834	-0.1473	-0.1396	0.0645	0.2350	0.3420	0.6505

The table shows the performance of the benchmark model as well as the worst, median, best performing model. A test that ignores the full space of models, and test the significance of the best model, relative to the benchmark would yield the naive “ p -value”. The SPA_c p -values controls for the full model space. The SPA_l and SPA_u provide a lower and upper bound for the true p -values respectively, whereas the SPA_c p -values are consistent for the true p -values.

Table 4: IBM Data

Benchmark: ARCH(1)								
Criterion	Performance				p -values			
	Bench.	Worst	Median	Best	Naive	SPA _l	SPA _c	SPA _u
MSE ₂	-30.929	-31.028	-24.968	-22.160	0.0060	0.0005	0.0010	0.0010
MSE ₁	-0.8047	-0.8108	-0.6222	-0.5599	0.0055	0.0000	0.0000	0.0000
QLIKE	-2.9177	-2.9237	-2.7665	-2.7423	0.0005	0.0000	0.0000	0.0000
R ² LOG	-0.4837	-0.5357	-0.4015	-0.3776	0.0150	0.0015	0.0015	0.0015
MAD ₂	-3.0774	-3.5636	-2.9847	-2.8111	0.0020	0.0095	0.0100	0.0115
MAD ₁	-0.6191	-0.7092	-0.5915	-0.5552	0.0015	0.0075	0.0085	0.0095

Benchmark: GARCH(1, 1)								
Criterion	Performance				p -values			
	Bench.	Worst	Median	Best	Naive	SPA _l	SPA _c	SPA _u
MSE ₂	-25.232	-31.028	-24.968	-22.160	0.0385	0.0300	0.0320	0.0340
MSE ₁	-0.6317	-0.8108	-0.6222	-0.5599	0.0330	0.0260	0.0315	0.0375
QLIKE	-2.7711	-2.9237	-2.7665	-2.7423	0.0235	0.0405	0.0490	0.0550
R ² LOG	-0.4086	-0.5357	-0.4015	-0.3776	0.0205	0.0655	0.0735	0.0870
MAD ₂	-3.0307	-3.5636	-2.9847	-2.8111	0.0020	0.0080	0.0095	0.0100
MAD ₁	-0.6018	-0.7092	-0.5915	-0.5552	0.0025	0.0045	0.0045	0.0055

The table shows the performance of the benchmark model as well as the worst, median, best performing model. A test that ignores the full space of models, and test the significance of the best model, relative to the benchmark would yield the naive “ p -value”. The SPA_c p -values controls for the full model space. The SPA_l and SPA_u provide a lower and upper bound for the true p -values respectively, whereas the SPA_c p -values are consistent for the true p -values.

Table 5: IBM Data (Reality Check)

Benchmark: ARCH(1)								
Criterion	Performance				p -values			
	Bench.	Worst	Median	Best	Naive	SPA_t^o	SPA_c^o	RC
MSE ₂	-30.929	-31.029	-24.977	-22.161	0.0060	0.0200	0.0200	0.0200
MSE ₁	-0.8047	-0.8108	-0.6222	-0.5599	0.0055	0.0200	0.0200	0.0200
QLIKE	-2.9177	-2.9237	-2.7670	-2.7423	0.0005	0.0050	0.0050	0.0050
R ² LOG	-0.4837	-0.5357	-0.4016	-0.3776	0.0150	0.0580	0.0710	0.0710
MAD ₂	-3.0774	-3.5636	-2.9850	-2.8111	0.0020	0.1195	0.1600	0.1915
MAD ₁	-0.6191	-0.7092	-0.5915	-0.5552	0.0015	0.0950	0.1265	0.1440

Benchmark: GARCH(1, 1)								
Criterion	Performance				p -values			
	Bench.	Worst	Median	Best	Naive	SPA_t^o	SPA_c^o	RC
MSE ₂	-25.232	-31.029	-24.977	-22.161	0.0385	0.1105	0.1110	0.1610
MSE ₁	-0.6317	-0.8108	-0.6222	-0.5599	0.0330	0.1000	0.1495	0.2810
QLIKE	-2.7711	-2.9237	-2.7670	-2.7423	0.0235	0.0940	0.1205	0.3725
R ² LOG	-0.4086	-0.5357	-0.4016	-0.3776	0.0205	0.2880	0.3485	0.5955
MAD ₂	-3.0307	-3.5636	-2.9850	-2.8111	0.0020	0.0735	0.1140	0.1715
MAD ₁	-0.6018	-0.7092	-0.5915	-0.5552	0.0025	0.0575	0.1180	0.1515

The table shows how a test statistic of the RC (supremum of unscaled relative performance) is unable to detect the inferiority of the GARCH(1,1). Compare the p -values, to those in Table 4, which are based on the more powerful SPA-test.

Table 6: Summary Statistics for the In-Sample Evaluation.

Criterion	Method for Estimating Realized volatility						
	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	0.0330	0.0250	0.0170	0.0200	0.0220	0.0120	0.0115
MSE ₂	0.0310	0.0280	0.0195	0.0185	0.0200	0.0155	0.0150
QLIKE	0.0540	0.0410	0.0220	0.0240	0.0225	0.0120	0.0115
R ² LOG	0.0615	0.0940	0.0385	0.0330	0.0390	0.0340	0.0285
MAD ₁	0.0080	0.0765	0.0710	0.0560	0.0620	0.0730	0.0790
MAD ₂	0.0055	0.0540	0.0470	0.0700	0.0580	0.0955	0.0830

This table reports the p -values of the SPA_c-test using the GARCH(1,1), in the analysis of IBM returns. The different p -values correspond to different methods to estimate volatilities, using intra-day returns. As can be seen, the evidence against the GARCH(1,1) being superior, is fairly robust to the method of intra-day estimation of volatility.

A FORECAST COMPARISON OF VOLATILITY MODELS

Table 7: Models with Gaussian error distribution and mean zero

Model	Exchange Rate Data						IBM Data						Mean
	L ₁	L ₂	L ₃	L ₄	L ₅	L ₆	L ₁	L ₂	L ₃	L ₄	L ₅	L ₆	
ARCH(1)	4.6	4.0	0.9	4.3	8.2	5.5	1.5	1.5	1.5	8.2	28.6	20.7	7.4
GARCH(1, 1)	86.0	93.6	97.6	93.3	89.7	90.9	42.6	40.7	42.2	36.8	35.0	35.3	65.3
GARCH(2,1)	84.5	87.5	94.8	86.6	86.6	89.1	25.5	15.2	28.6	12.5	14.3	12.8	53.2
GARCH(1,2)	85.1	80.5	65.3	87.5	91.8	91.8	40.7	30.7	41.6	25.2	27.7	28.9	58.1
GARCH(2,2)	89.1	88.4	72.6	92.1	96.4	96.4	43.2	29.2	42.6	18.2	23.7	21.6	59.4
IGARCH(1, 1)	7.3	7.9	17.9	8.8	6.1	8.2	13.1	3.6	14.6	1.8	2.4	2.1	7.8
IGARCH(2,1)	6.7	7.6	17.0	8.5	5.8	7.9	15.8	7.0	17.0	5.2	5.5	5.5	9.1
IGARCH(1,2)	4.0	6.1	14.6	7.9	4.6	6.4	13.7	4.3	14.9	2.1	3.0	2.7	7.0
IGARCH(2,2)	10.6	8.8	14.0	6.1	7.9	8.8	37.1	8.8	31.9	7.3	7.0	6.4	12.9
TS-GARCH(1, 1)	54.4	58.7	73.6	61.1	35.3	40.1	86.3	68.7	84.5	29.5	24.0	24.3	53.4
TS-GARCH(2,1)	57.1	57.4	72.3	58.4	32.5	38.3	72.6	68.1	82.7	35.3	31.3	31.6	53.1
TS-GARCH(1,2)	91.8	88.8	86.6	84.2	76.6	72.6	87.2	71.1	90.3	33.1	26.7	27.1	69.7
TS-GARCH(2,2)	94.8	95.7	91.2	93.0	82.4	79.9	79.3	69.9	85.7	35.9	31.6	31.0	72.5
A-GARCH(1, 1)	71.7	78.1	86.0	79.6	79.6	82.7	47.1	65.0	49.5	81.5	70.5	72.3	72.0
A-GARCH(2,1)	60.5	59.9	54.1	62.3	67.5	74.8	36.5	38.6	30.7	66.6	58.4	62.3	56.0
A-GARCH(1,2)	85.7	81.2	66.3	86.9	92.7	92.7	45.0	63.8	46.5	86.0	74.8	80.2	75.2
A-GARCH(2,2)	20.1	19.5	8.5	18.8	40.1	41.0	31.6	35.6	29.8	62.9	56.5	59.0	35.3
NA-GARCH(1, 1)	56.5	68.7	75.7	72.9	71.7	77.2	49.5	59.9	49.8	73.6	64.1	67.2	65.6
NA-GARCH(2,1)	47.1	51.7	50.2	54.1	54.7	60.5	27.7	29.8	20.4	45.3	41.0	51.4	44.5
NA-GARCH(1,2)	87.5	82.1	69.3	84.5	93.0	92.1	48.6	57.8	45.6	74.8	67.5	70.5	72.8
NA-GARCH(2,2)	8.8	9.1	2.1	10.0	16.1	17.3	29.2	28.6	20.1	44.7	37.7	48.0	22.6
V-GARCH(1, 1)	31.9	40.7	36.8	39.5	80.9	71.7	8.5	24.9	9.1	90.9	99.4	99.4	52.8
V-GARCH(2,1)	31.3	29.2	24.9	27.1	70.8	55.0	4.3	14.0	4.3	51.7	82.7	80.5	39.6
V-GARCH(1,2)	28.0	36.5	24.0	45.6	84.5	77.5	9.1	24.0	8.5	89.4	99.1	99.1	52.1
V-GARCH(2,2)	18.2	15.8	10.0	13.1	44.1	35.0	3.3	12.8	3.3	46.2	78.7	76.9	29.8
THR-GARCH(1, 1)	25.2	27.7	37.4	35.6	19.8	23.1	69.9	60.5	64.4	28.6	22.8	28.3	36.9
THR-GARCH(2,1)	24.3	25.2	30.4	29.5	14.0	15.8	55.9	41.9	58.7	25.8	18.5	22.2	30.2
THR-GARCH(1,2)	91.2	86.6	84.2	80.9	66.9	62.3	71.4	61.1	72.6	28.0	22.2	28.0	62.9
THR-GARCH(2,2)	8.5	11.2	10.9	14.0	10.0	10.9	55.6	41.6	58.4	25.5	18.2	21.9	23.9
GJR-GARCH(1, 1)	79.0	89.7	95.7	91.2	84.8	88.4	26.1	23.7	28.3	41.3	52.9	63.2	63.7
GJR-GARCH(2,1)	69.3	75.1	83.9	79.0	77.5	81.2	18.8	19.1	16.1	35.6	48.6	58.7	55.2
GJR-GARCH(1,2)	83.6	78.7	58.7	82.7	90.6	90.0	24.6	20.7	24.0	39.2	55.0	64.7	59.4
GJR-GARCH(2,2)	15.2	17.9	16.1	20.7	41.3	44.4	49.8	32.8	50.5	21.9	28.9	30.4	30.8
LOG-GARCH(1, 1)	81.2	72.0	79.0	65.0	52.0	51.7	82.1	77.8	81.8	43.2	36.2	34.3	63.0
LOG-GARCH(2,1)	84.2	69.6	76.3	59.6	51.4	51.1	63.2	47.7	67.5	20.4	16.4	16.1	52.0
LOG-GARCH(1,2)	99.4	98.5	93.9	99.4	97.6	97.3	79.0	73.6	81.2	38.6	33.1	31.9	77.0
LOG-GARCH(2,2)	100.0	100.0	95.1	99.7	99.4	99.1	62.9	42.2	62.0	17.9	13.4	13.7	67.1
EGARCH(1, 1)	37.1	38.6	40.7	38.6	38.6	36.5	70.8	76.6	63.2	60.2	56.8	55.3	51.1
EGARCH(2,1)	42.9	39.5	41.6	38.3	35.6	33.7	53.2	50.5	52.9	38.9	35.3	39.8	41.8
EGARCH(1,2)	99.7	99.7	95.4	98.5	97.9	97.6	68.7	73.3	69.6	55.0	52.6	53.2	80.1
EGARCH(2,2)	11.6	13.4	13.1	15.5	14.9	15.2	52.9	48.0	51.4	37.7	34.7	38.6	28.9
NGARCH(1, 1)	83.0	91.2	97.9	92.4	72.3	79.0	96.7	67.5	90.9	21.0	17.6	15.5	68.7
NGARCH(2,1)	80.2	81.8	96.4	83.9	64.7	74.5	83.6	34.0	77.5	11.9	9.1	9.1	58.9
NGARCH(1,2)	92.7	94.8	88.8	95.1	88.4	87.2	97.3	67.2	92.7	19.5	14.9	14.6	71.1
NGARCH(2,2)	94.5	96.4	93.0	98.8	92.1	91.5	83.9	33.1	74.5	11.6	8.8	8.8	65.6
A-PARCH(1, 1)	43.8	60.8	76.6	71.4	46.5	51.4	81.2	53.2	69.3	21.6	15.5	16.7	50.7
A-PARCH(2,1)	38.3	48.6	58.1	58.1	37.7	43.5	56.2	31.6	57.8	17.3	12.5	13.1	39.4
A-PARCH(1,2)	93.0	95.4	89.7	95.4	89.1	87.8	84.5	55.0	76.3	22.5	16.1	17.0	68.5
A-PARCH(2,2)	52.0	65.0	52.9	75.7	63.8	66.3	56.5	31.9	58.1	17.6	12.2	13.4	47.1
GQ-ARCH(1, 1)	71.4	77.8	86.3	79.9	79.3	83.0	47.4	65.7	49.2	81.8	70.8	72.6	72.1
GQ-ARCH(2,1)	77.8	95.1	99.7	97.0	80.5	91.2	18.2	27.7	13.1	48.0	48.0	52.3	62.4
GQ-ARCH(1,2)	85.4	80.9	66.0	87.2	92.4	92.4	45.3	64.1	46.8	85.7	74.5	79.9	75.1
GQ-ARCH(2,2)	21.3	22.2	21.6	25.2	27.1	27.4	9.7	8.2	9.4	9.7	12.8	11.6	17.2
H-GARCH(1, 1)	39.5	48.0	54.7	49.2	44.7	46.8	67.8	18.5	56.5	10.6	8.2	8.2	37.7
AUG-GARCH(1, 1)	43.5	46.8	47.7	44.7	50.8	48.9	58.1	12.2	51.1	9.1	6.4	7.0	35.5

Relative performance ranking. Each row corresponds to a particular model, and a score shows the percentage of models (out of the total of 330) that performed worse than the particular model, measured in terms of a given loss function. Thus, the worst, median, and best models score 0, 50, and 100 respectively. The loss functions, given in (5), . . . , (10), are here denoted by L₁, . . . , L₆. The last column is the average of the 12 scores.

Table 8: Models with t -distributed errors and zero mean

Model	Exchange Rate Data						IBM Data						Mean
	L ₁	L ₂	L ₃	L ₄	L ₅	L ₆	L ₁	L ₂	L ₃	L ₄	L ₅	L ₆	
ARCH(1)	3.3	1.8	0.0	0.6	6.7	3.6	0.3	0.3	0.6	6.4	24.9	17.6	5.5
GARCH(1, 1)	77.2	73.9	83.3	72.0	69.0	71.1	19.1	23.4	20.7	31.0	40.1	36.8	51.5
GARCH(2, 1)	79.6	73.3	82.4	69.6	74.5	75.7	19.5	20.4	23.7	27.7	37.4	35.6	51.6
GARCH(1, 2)	72.0	62.0	53.2	60.5	71.4	69.9	19.8	25.2	21.3	31.3	39.2	36.5	46.9
GARCH(2, 2)	80.9	75.4	63.8	75.4	85.4	84.8	28.6	30.1	27.7	30.7	39.5	37.1	54.9
IGARCH(1, 1)	2.4	3.6	6.4	3.3	1.8	2.1	9.4	1.8	11.9	0.0	0.0	0.0	3.6
IGARCH(2, 1)	2.1	3.3	6.7	3.6	2.4	2.7	12.2	4.0	15.2	2.4	1.8	1.8	4.9
IGARCH(1, 2)	1.8	3.0	5.8	3.0	2.1	2.4	10.6	2.4	12.5	0.9	0.6	0.6	3.8
IGARCH(2, 2)	7.9	5.2	7.9	5.2	3.3	3.3	16.4	6.4	21.0	4.6	3.6	4.3	7.4
TS-GARCH(1, 1)	34.7	32.5	41.0	34.7	15.2	15.5	78.4	80.2	83.0	53.8	47.7	45.3	46.8
TS-GARCH(2, 1)	35.9	34.0	43.5	35.3	17.3	18.2	74.2	78.1	86.9	48.3	42.6	39.2	46.1
TS-GARCH(1, 2)	47.7	44.7	47.4	44.1	30.4	30.1	79.6	81.5	87.8	56.5	45.9	42.6	53.2
TS-GARCH(2, 2)	59.6	67.2	71.4	71.7	47.7	50.8	75.1	78.4	87.2	48.6	42.9	39.5	61.7
A-GARCH(1, 1)	72.9	70.2	78.1	67.5	66.0	67.5	34.7	49.8	33.4	84.2	86.0	86.6	66.4
A-GARCH(2, 1)	60.8	54.7	49.2	48.6	58.1	59.9	34.3	42.9	38.6	55.6	62.6	57.8	51.9
A-GARCH(1, 2)	58.4	50.2	41.3	55.3	62.3	64.4	33.1	47.1	36.5	77.2	77.5	77.8	56.8
A-GARCH(2, 2)	19.5	18.5	11.6	16.1	31.6	31.3	52.3	71.7	56.2	77.8	83.6	76.6	45.6
NA-GARCH(1, 1)	66.9	66.9	72.0	64.7	60.5	62.6	41.3	58.7	39.5	96.0	91.2	93.0	67.8
NA-GARCH(2, 1)	56.2	51.4	49.5	47.7	53.8	56.5	41.6	59.0	40.4	95.7	90.6	92.4	61.2
NA-GARCH(1, 2)	61.1	55.6	44.1	60.8	69.6	70.2	39.2	55.6	41.3	93.9	88.1	89.1	64.1
NA-GARCH(2, 2)	12.8	12.5	7.0	11.6	21.9	24.3	38.3	54.1	41.0	87.8	86.9	87.8	40.5
V-GARCH(1, 1)	29.8	34.7	30.1	30.4	67.2	52.9	6.1	18.2	6.7	72.6	97.9	97.9	45.4
V-GARCH(2, 1)	34.0	34.3	27.1	30.1	77.2	57.8	1.8	11.2	1.8	48.9	93.3	91.5	42.4
V-GARCH(1, 2)	22.2	21.0	20.4	19.8	41.0	28.6	5.2	17.3	5.5	71.1	97.3	97.3	37.2
V-GARCH(2, 2)	13.7	10.0	4.3	9.1	21.0	12.5	11.9	26.4	10.0	68.4	94.8	94.8	31.4
THR-GARCH(1, 1)	23.7	24.0	27.7	24.0	11.9	11.6	77.2	88.8	74.2	88.4	75.4	74.2	50.1
THR-GARCH(2, 1)	23.1	23.1	25.2	21.9	9.7	9.4	82.4	83.6	94.2	63.2	48.3	50.2	44.5
THR-GARCH(1, 2)	48.0	46.2	47.1	45.3	38.0	39.2	74.8	89.4	79.0	90.0	75.7	75.7	62.4
THR-GARCH(2, 2)	12.5	14.3	19.5	17.0	11.2	13.4	99.4	99.7	99.4	92.7	83.3	76.3	53.2
GJR-GARCH(1, 1)	67.8	69.9	78.4	70.2	65.0	67.8	24.0	36.5	26.1	68.1	80.2	84.2	61.5
GJR-GARCH(2, 1)	63.2	61.1	62.3	59.9	59.6	64.1	28.0	39.2	34.7	51.1	62.0	58.4	53.6
GJR-GARCH(1, 2)	47.4	41.0	33.7	43.8	56.5	58.1	22.5	37.7	27.4	69.0	79.9	83.9	50.1
GJR-GARCH(2, 2)	16.7	17.0	18.8	16.7	31.9	36.2	51.1	66.3	51.7	64.4	65.3	66.3	41.9
LOG-GARCH(1, 1)	55.9	42.2	44.7	36.2	29.8	25.5	66.6	82.4	73.9	71.7	67.2	61.4	54.8
LOG-GARCH(2, 1)	66.0	45.6	48.0	37.4	33.4	30.7	59.9	74.2	65.3	61.1	60.5	54.7	53.1
LOG-GARCH(1, 2)	96.4	87.2	67.8	71.1	82.7	72.3	64.7	81.2	72.9	72.3	64.7	57.4	74.2
LOG-GARCH(2, 2)	97.9	92.7	81.5	79.3	90.0	79.6	60.5	73.9	65.0	60.5	60.2	54.4	74.6
EGARCH(1, 1)	35.6	30.7	33.4	28.6	24.0	21.6	65.0	90.6	61.7	99.1	95.7	95.7	56.8
EGARCH(2, 1)	40.1	31.6	31.0	25.8	23.7	21.3	61.1	83.9	71.4	92.1	84.2	83.6	54.2
EGARCH(1, 2)	80.5	55.3	35.6	47.1	59.9	49.8	62.6	86.3	62.6	97.0	94.2	93.9	68.7
EGARCH(2, 2)	15.8	14.9	18.2	14.9	15.5	14.9	97.9	98.5	98.8	97.6	91.8	88.4	55.6
NGARCH(1, 1)	51.4	53.8	62.9	53.2	32.8	36.8	93.3	90.9	93.6	54.4	53.5	47.7	60.4
NGARCH(2, 1)	55.0	55.0	65.7	52.3	36.2	40.4	93.6	87.8	96.7	50.8	46.8	45.0	60.4
NGARCH(1, 2)	58.7	55.9	55.3	55.0	42.2	42.6	95.1	93.9	96.0	57.8	48.9	45.9	62.3
NGARCH(2, 2)	65.7	71.4	76.9	75.1	51.1	53.8	92.1	87.5	97.0	51.4	46.5	44.4	67.7
A-PARCH(1, 1)	35.0	37.7	48.6	41.9	24.6	25.2	88.8	95.7	79.9	81.2	72.9	69.9	58.5
A-PARCH(2, 1)	29.2	28.3	37.1	33.7	18.8	20.7	92.4	97.6	97.3	76.3	68.7	67.5	55.6
A-PARCH(1, 2)	46.2	46.5	44.4	48.9	45.6	44.7	89.4	97.0	91.5	82.7	72.6	69.6	64.9
A-PARCH(2, 2)	25.8	28.9	31.9	39.2	27.4	31.0	100.0	100.0	100.0	96.7	87.8	83.0	62.6
GQ-ARCH(1, 1)	73.6	70.8	78.7	68.1	66.3	68.1	35.0	50.2	33.7	84.5	86.3	86.9	66.8
GQ-ARCH(2, 1)	74.8	88.1	98.8	90.3	71.1	83.9	6.7	9.4	4.9	16.4	29.8	26.4	50.1
GQ-ARCH(1, 2)	59.3	50.8	43.2	56.5	63.2	64.7	32.8	47.4	36.2	77.5	77.8	78.1	57.3
GQ-ARCH(2, 2)	18.5	20.4	21.0	21.3	17.0	20.4	15.5	13.4	10.6	16.1	26.4	22.5	18.6
H-GARCH(1, 1)	76.9	79.9	90.0	81.8	59.0	59.3	96.0	81.8	69.9	46.8	45.6	45.6	69.4
AUG-GARCH(1, 1)	45.3	51.1	61.1	54.7	34.3	38.9	95.4	84.5	78.4	49.5	42.2	41.6	56.4

Relative performance ranking. Each row corresponds to a particular model, and a score shows the percentage of models (out of the total of 330) that performed worse than the particular model, measured in terms of a given loss function. Thus, the worst, median, and best models score 0, 50, and 100 respectively. The loss functions, given in (5), . . . , (10), are here denoted by L₁, . . . , L₆. The last column is the average of the 12 scores.

A FORECAST COMPARISON OF VOLATILITY MODELS

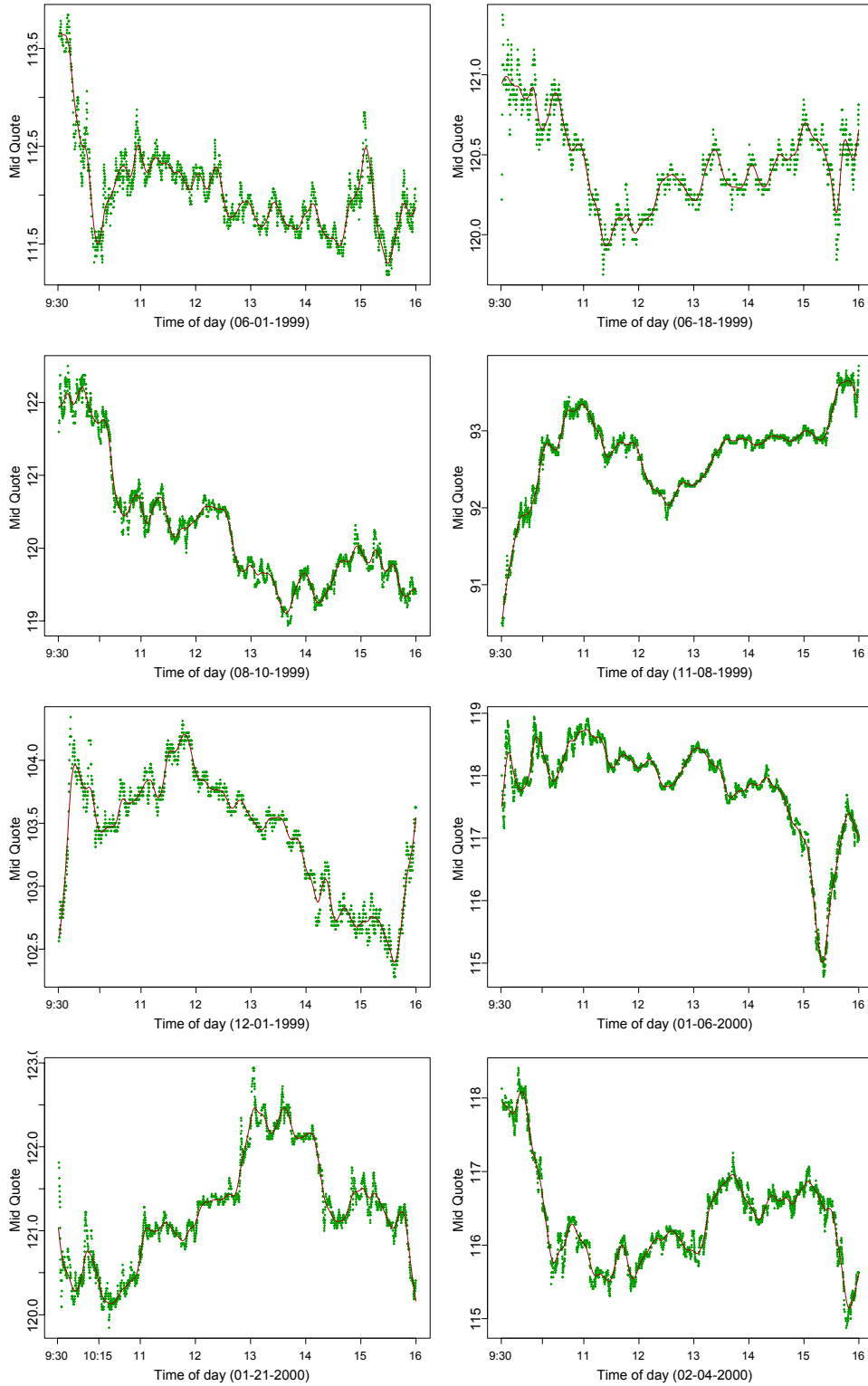


Figure 1: Intra-day mid quotes, and fitted spline-curves.

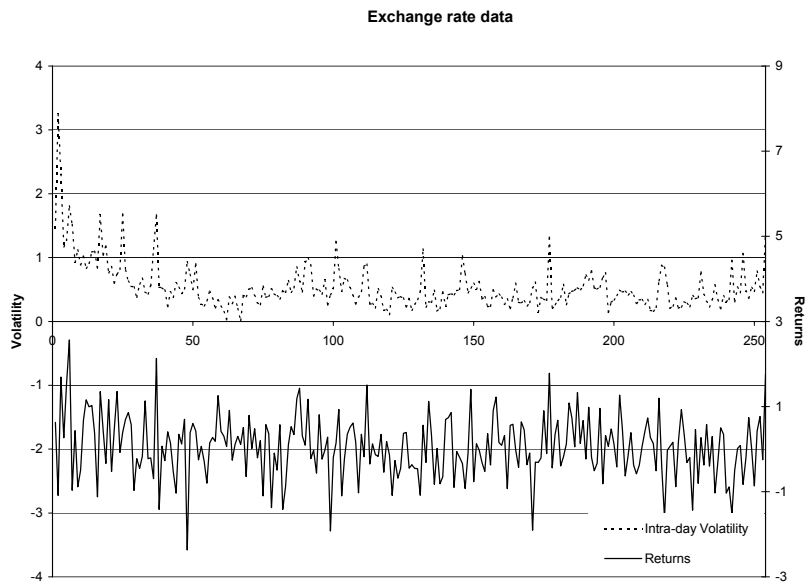


Figure 2: The intra-day volatility and returns of the DM-\$ exchange rate data.

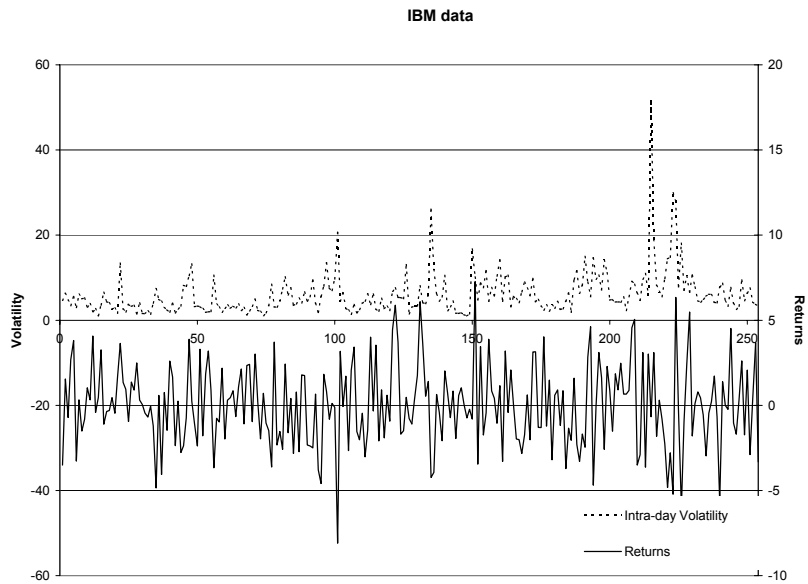


Figure 3: The intra-day volatility and returns of the DM-\$ IBM data.

A FORECAST COMPARISON OF VOLATILITY MODELS

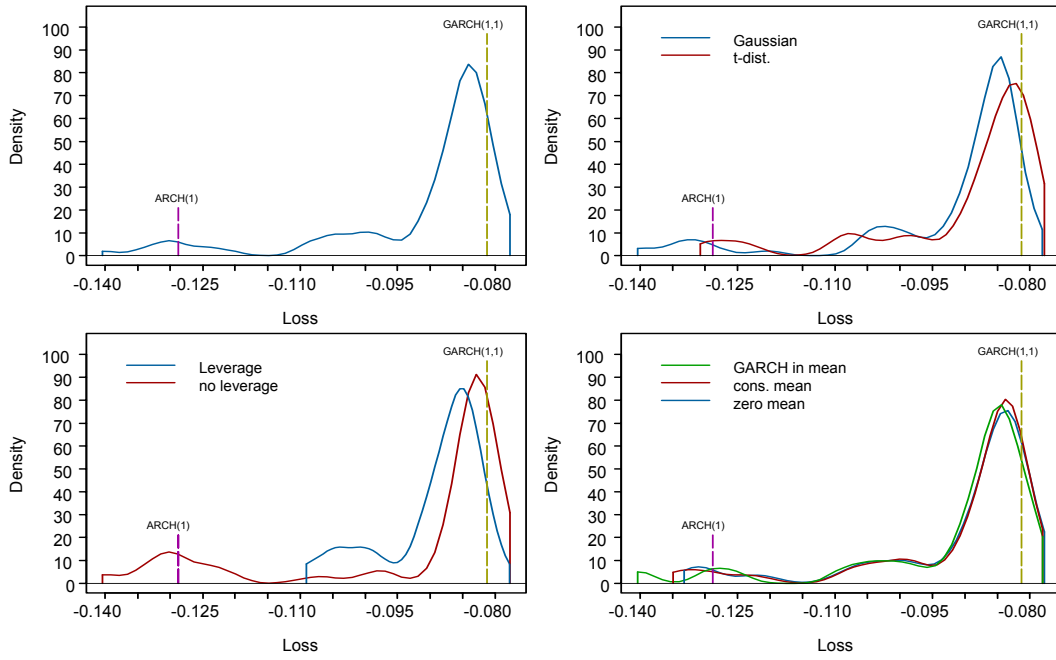


Figure 4: Population of Average Model Forecasts: Exchange Rate Data and MSE_2 Loss Function.

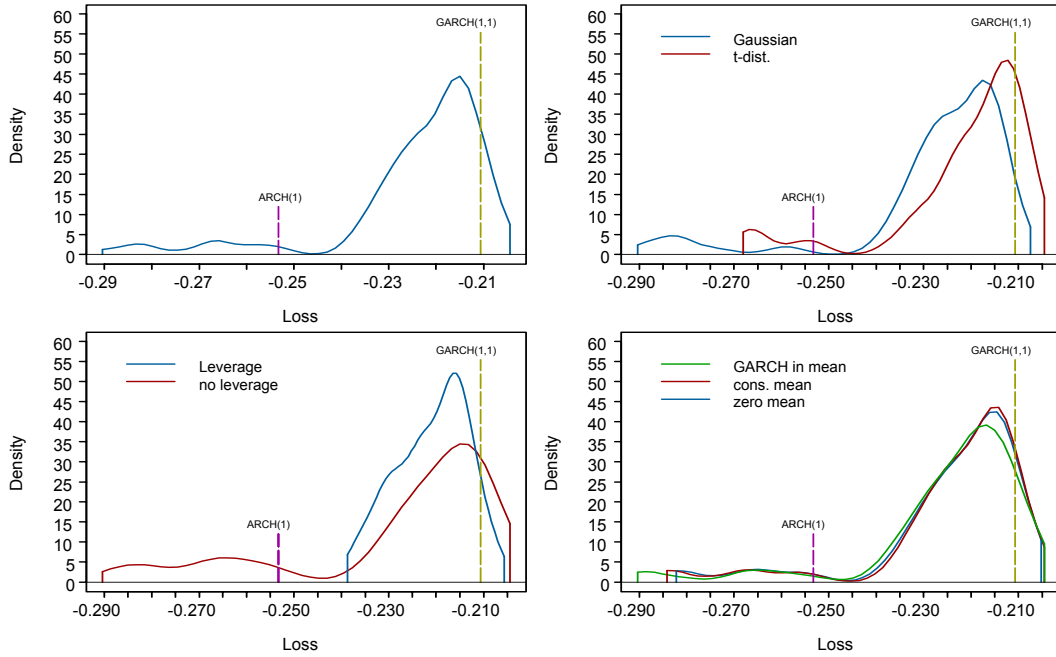


Figure 5: Population of Average Model Forecasts: Exchange Rate Data and MAD_2 Loss Function.

A FORECAST COMPARISON OF VOLATILITY MODELS

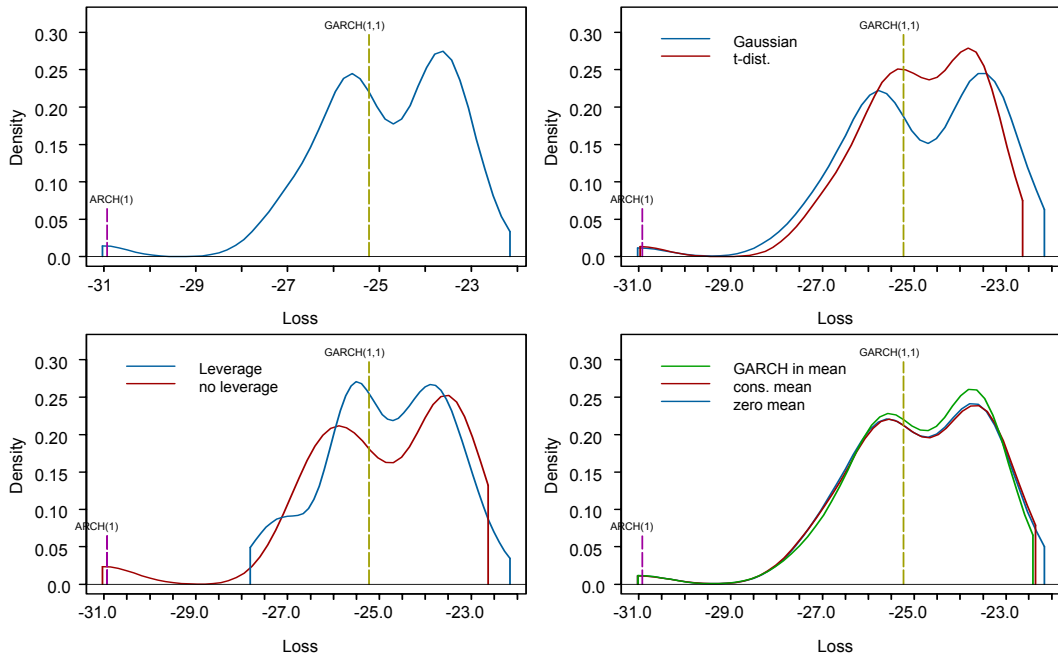


Figure 6: Population of Average Model Forecasts: IBM Data and MSE_2 Loss Function.

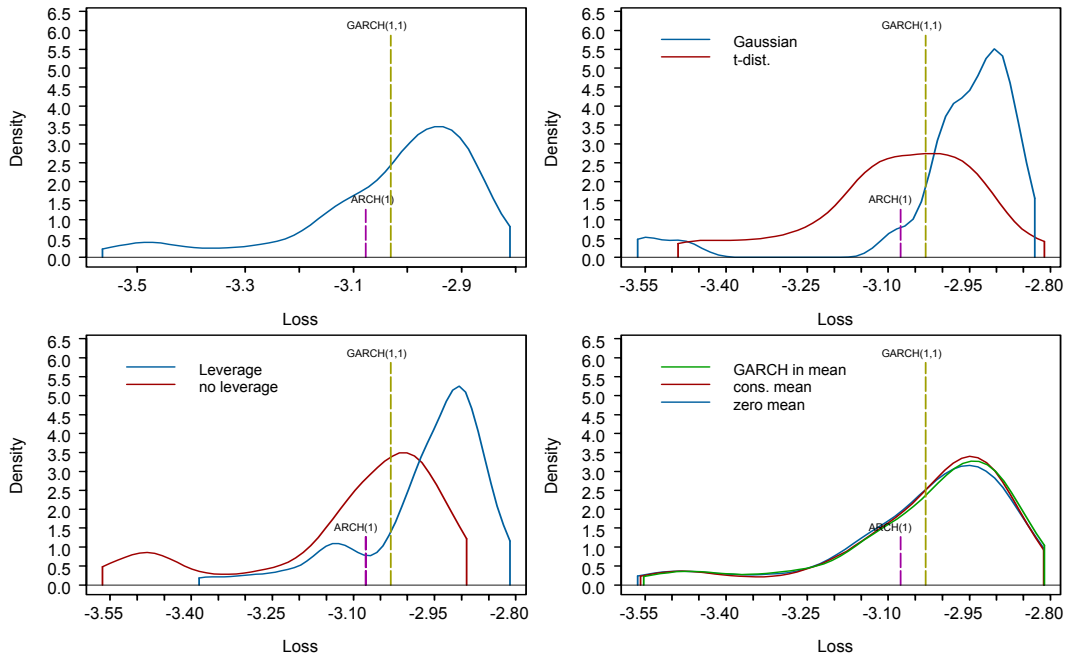


Figure 7: Population of Average Model Forecasts: IBM Data and MAD_2 Loss Function.