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


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

Asger Lunde
Aalborg University, Economics
Fibirgerstraede 3
DK 9220 Aalborg Ø
Phone: (+45) 9635-8176
Email: alunde@cls.dk


March 8, 2001


**Abstract**

By using intra-day returns to calculate a measure for the time-varying volatility, Andersen and Bollerslev (1998a) established that volatility models do provide good forecasts of the conditional variance.

In this paper, we take the same approach and use intra-day estimated measures of volatility to compare volatility models. Our objective is to evaluate whether the evolution of volatility models has led to better forecasts of volatility when compared to the first "species" of volatility models.

We make an out-of-sample comparison of 330 different volatility models using daily exchange rate data (DM/$) and IBM stock prices. Our analysis does not point to a single winner amongst the different volatility models, as it is different models that are best at forecasting the volatility of the two types of assets. Interestingly, the best models do not provide a significantly better forecast than the GARCH(1,1) model. This result is established by the tests for superior predictive ability of White (2000) and Hansen (2001). If an ARCH(1) model is selected as the benchmark, it is clearly outperformed.



We thank Tim Bollerslev for providing us with the exchange rate data set, and Sivan Ritz for suggesting numerous clarifications. All errors remain our responsibility.

# 1 Introduction

Time-variation in 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 such models have been developed, starting with the ARCH model of Engle (1982).

The fact that the conditional variance is unobserved has affected the development of volatility models and has made it difficult to evaluate and compare the different models. Therefore the models with poor forecasting abilities have not been identified, and this may explain why so many models have been able to coexist. In addition, there does not seem to be a natural and intuitive way to model conditional heteroskedasticity – different models attempt to capture different features that are thought to be important. For example, some models allow the volatility to react asymmetrically to positive and negative changes in returns. Features of this kind are typically found to be very significant in in-sample analyses. However, the significance may be a result of a misspecification, and it is therefore not certain that the models with such features result in better out-of-sample forecasts, compared to the forecasts of more parsimonious models.

When evaluating the performance of a volatility model, the unobserved variance was often substituted with squared returns, and this commonly led to a very poor out-of-sample performance. The poor out-of-sample performance instigated a discussion of the practical relevance of these models, which was resolved by Andersen and Bollerslev (1998a). Rather than using squared inter-day returns, which are very noisy measures of daily volatility, Andersen and Bollerslev based their evaluation on an estimated measure of the volatility using intra-day returns, which resulted in a good out-of-sample performance of volatility models. This indicates that the previously found poor performance can be explained by the use of a noisy measure of the volatility.

In this paper, we compare volatility models using an intra-day estimate measures of realized volatility. Since this precise measures of volatility makes it easier to evaluate the performance of the individual models, it also becomes easier to compare different models. If some models are better than others in terms of their predictive ability, then it should be easier to determine this superiority, because the noise in the evaluation is reduced. We evaluate the relative performance

of the various volatility models in terms of their predictive ability of realized volatility, by using the recently developed tests for superior predictive ability of White (2000) and Hansen (2001). These tests are also referred to as tests for data snooping. Unfortunately, it is not clear which criteria one should use to compare the models, as was pointed out by Bollerslev, Engle, and Nelson (1994) and Diebold and Lopez (1996). Therefore, we use seven different criteria for our comparison, which include standard criteria such as the mean squared error (MSE) criterion, a likelihood criterion, and the mean absolute deviation criterion, which is less sensitive to extreme mispredictions, compared to the MSE.

Given a benchmark model and an evaluation criterion, the tests for data snooping enable us to test whether any of the competing models are significantly better than the benchmark. We specify two different benchmark models. An ARCH(1) model and a GARCH(1,1) model. The tests for data snooping clearly point to better models in the first case, but the GARCH(1,1) is not significantly outperformed in the data sets we consider. Although the analysis in one of the data sets does point to the existence of a better model than the GARCH(1,1) when using the mean squared forecast error as the criterion, this result does not hold up to other criteria that are more robust to outliers, such as the mean absolute deviation criterion.

The power properties of tests for data snooping can, in some applications, be poor. But our rejection of the ARCH(1) indicates that this is not a severe problem in this analysis. The fact that the tests for data snooping are not uncritical to any choice of benchmark is comforting.

This paper is organized as follows. Section 2 describes the universe of volatility models that we include in the analysis. It also describes the estimation of the models. Section 3 describes the performance criteria and the data we use to compare the models. Section 4 describes the tests for data snooping. Section 5 contains our results and Section 6 contains concluding remarks.

## 2 The GARCH Universe

We use the notation of Hansen (1994) to set up 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 $\sigma$-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 asset prices, $p_t$. We

define the compounded return by

$$r_t = \log(p_t) - \log(p_{t-1}), \qquad t = -R+1, \ldots, n, \tag{1}$$

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, $t = -R+1, \ldots, 0$, and an evaluation period with $n$ periods, $t = 1, \ldots, n$.

Our objective is to model the conditional density or $r_t$, 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 exists). Subsequently we can define the standardized residuals, which are denoted by $e_t = (r_t - \mu_t)/\sigma_t$, $t = -R+1, \ldots, n$. We denote the conditional density function of the standardized residuals 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 one of $e_t$ by the following relationship

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

Thus, a 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. Which make 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 $\theta$ 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 $\theta$, we require that $\psi_t$ is observable[1] at time $t-1$. This yields a complete specification of the conditional distribution of $r_t$.

As described above, we can divide the vector of time varying parameters into three components,

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

where $\mu_t$ is the conditional mean (the *location* parameter), $\sigma_t$ is the conditional standard deviation (the *scale* parameter), and $\eta_t$ are the remaining (*shape*) parameters of the conditional

[1] This 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, $\mu_t$, is given by

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

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

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

In financial time-series, it is often important to model the distribution with a higher precision than the first two moments. This is achieved through a modelling of the density function for the standardized residuals, $e_t$, through the shape parameters $\eta_t$.

Most of the existing GARCH-type models can be expressed in this framework, and when expressed in this framework, the corresponding $\eta_t$'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 $\eta_t$ constant, but we hope to relax this restrictive assumption in future research. Models with non-constant $\eta_t$ include Hansen (1994) and Harvey and Siddique (1999). 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 and Tauchen (1989).

## 2.1 The Conditional Mean

Our modelling of the conditional mean, $\mu_t$, takes the form

$$m_t = \mu_0 + \mu_1 \zeta(\sigma_{t-1})$$

where $\zeta(x) = x^2$. The three specifications we include in the analysis are: the GARCH-in-mean suggested by Engle, Lillen, and 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. Our aim was to include all parametric specifications that have been suggested in the literature. But as stated earlier we restrict our analysis to parametric specifications, specifically the parameterizations given in Table 2. The

specifications for $\sigma_t$, that we included in our analysis are 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-GARCH[2], the NA-GARCH and the V-GARCH models suggested by Engle and Ng (1993), the threshold GARCH model (Thr.-GARCH) by Zakoian (1994), the GJR-GARCH model of Glosten, Jagannathan, and Runkle (1993), the log-ARCH by Geweke (1986) and Pantula (1986), the EGARCH, the NGARCH of Higgins and Bera (1992), the A-PARCH model proposed in Ding, Granger, and Engle (1993), the GQ-ARCH suggested by Sentana (1995), the H-GARCH of Hentshel (1995), and finally the Aug-GARCH suggested by Duan (1997).

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 includes several of the other models as special cases.

The Aug-GARCH model has not (to our knowledge) been applied in published work. Nevertheless, we include it in our analysis, because 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 some sense and appear in published work. Finally, we include . Although, this results in a very large number of different volatility models, we have by no means exhausted the space of possible ARCH type model.

Given a particular volatility model, one can plot of $\sigma_t^2$ against $\varepsilon_{t-1}$, 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 and Schwert (1990), and later named the *News Impact Curve* by Engle and Ng (1993). The News Impact Curve, provides an easy way to interpret some aspects of the different volatility specifications and several of the models included in our analysis were compared using this method by Hentshel (1995).

The evolution of volatility models has been motivated by empirical findings and economic

[2]At 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 and Ng (1993) and rename the other models, e.g., the model by Hentshel (1995) is here called H-GARCH.

interpretations. Ding, Granger, and Engle (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 $\sigma_t$ rather than $\sigma_t^2$ or vice versa. More generally we can consider a modelling of $\sigma_t^\delta$ where $\delta$ is a parameter to be estimated. This is the motivation for the introduction of the *Box-Cox transformation* 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 suggests that stock returns are negatively correlated with changes in return 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. For further details on the leverage effect, see Engle and Patton (2000).

The specifications for the conditional variance, given in Table 2, contain parameters for the lag lengths, denoted by $p$ and $q$. In the present analysis we have included the four combinations of lag lengths $p, q = 1, 2$ for most models. The exceptions are the ARCH model where we only include $(p, q) = (1, 0)$ (the ARCH(1) model), and the H-GARCH and Aug-GARCH models, where we only include $(p, q) = (1, 1)$. The reason why we restrict our analysis to short and relatively few lag specification, is simply to keep the burden of estimation all the models at a manageable size. It is reasonable to expect that the models with more lag, will not result in more accurate forecasts than more parsimonious models. So to limit our attention to the models with short lags, should not affect our analysis.

### 2.3 The Density for the Standardized Returns

In the present analysis we only consider a Gaussian and a $t$-distributed specification for the density $g(e|\eta_t)$, the latter was first advocated by Bollerslev (1987). Thus, $\eta_t$ is held constant.

### 2.4 Estimation

The models are estimated using *inter*-day returns over the sample period $t = -R + 1, \ldots, 0$, whereas *intra*-day returns are used to construct a good estimate of the volatility. The intra-day estimated measures of volatilities are used to compare of the models, in the sample period $t = 1, \ldots, n$. The estimation is described in this subsection whereas the evaluation and comparison

are explained in Section 3.

All models were estimated using the method of maximum likelihood. The optimization problem was programmed in C++, and the likelihood functions were maximized using the simplex method described in Press, Teukolsky, Vetterling, and Flannary (1992). A total of 330 models were estimated[3].

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 our 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. However, given the large number of models and their complex nature, it is possible that one or more of the likelihood functions were not maximized. But we are comforted by the fact that we do not see any obvious inconsistencies across models. For example, for nested models we check 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 and Bollerslev (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.

# 3 Performance Metric

Given a forecast for volatility and a measure of realized volatility, it is non-trivial to evaluate the value of the forecast, as pointed out by Bollerslev, Engle, and Nelson (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 underlying the likelihood are dubious. Further, a good in-sample performance does not guarantee a good out-of-sample performance. This point is clearly relevant for our analysis. Most of the models we estimate have significant lags (that is $p$ or $q = 2$) in

[3]Due 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 refered to. The appendix can be downloaded from `http://www.socsci.auc.dk/~alunde`.

our in-sample analysis. But in the out-of-sample comparison, the models with more lags rarely perform better than the same model with fewer lags (measured by the $R^2$ of the regressions (3) and (4) below).

We index the $l$ volatility models by $k$, and denote model $k$'s forecast of $\sigma_t^2$ by $h_{k,t}^2$, $k = 1, \ldots, 330$ and $t = 1, \ldots, n$. The volatility models ability to make accurate predictions of the *realized 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. So the parameter estimates of $a$ and $b$ will primarily be determined by the observations where squared returns, $r_t^2$, have the largest values. This has been noted by Pagan and Schwert (1990) and Engle and Patton (2000)[4]. Therefore they advocate the regression

$$\log(r_t^2) = a + b\log(h_t^2) + u_t \tag{4}$$

which is less sensitive to "outliers", because severe mispredictions are given less weight than in (3).

In our analysis, we compare the models in terms of loss functions, some of which are even more robust to outliers. It is not possible to identify a unique and natural criterion for the comparison. So rather than making a single choice, we specify seven different loss functions,

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

which can be given different interpretations. The loss functions are

$$MSE_2 = n^{-1} \sum_{t=1}^{n} \left( \hat{\sigma}_t^2 - h_t^2 \right)^2 \tag{5}$$

$$MSE_1 = n^{-1} \sum_{t=1}^{n} \left( \hat{\sigma}_t - h_t \right)^2 \tag{6}$$

$$PSE = n^{-1} \sum_{t=1}^{n} \left( \hat{\sigma}_t^2 - h_t^2 \right)^2 h_t^{-4} \tag{7}$$

$$QLIKE = n^{-1} \sum_{t=1}^{n} \left( \log(h_t^2) + \hat{\sigma}_t^2 h_t^{-2} \right) \tag{8}$$

$$R2LOG = n^{-1} \sum_{t=1}^{n} \left[ \log(\hat{\sigma}_t^2 h_t^{-2}) \right]^2 \tag{9}$$

$$MAD_2 = n^{-1} \sum_{t=1}^{n} \left| \hat{\sigma}_t^2 - h_t^2 \right| \tag{10}$$

$$MAD_1 = n^{-1} \sum_{t=1}^{n} \left| \hat{\sigma}_t - h_t \right| \tag{11}$$

The criteria (5), (7), (8), and (9) were suggested by Bollerslev, Engle, and Nelson (1994), (here formulated in terms of a general estimated of volatility, $\hat{\sigma}_t$, rather than $\varepsilon_t^2$). The criteria (5) and (9) are (apart from the constant term, $a$) equivalent to using the $R^2$s from the regressions (3) and (4), respectively, the former is also known as the mean squared forecast error criterion. (7) measures the percentage squared errors, whereas (8) corresponds to the loss function implied by a Gaussian likelihood. The mean absolute deviation criteria (10) and (11) are interesting because they are more robust to outliers than, say, the mean squared forecast error criterion.

Estimation of volatility models usually results in highly significant in-sample 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 and Bollerslev (1998a) have since refuted this skepticism by demonstrating that well-specified volatility models do provide quite accurate forecasts of volatility. The problem is that $r_t^2$ is a noisy estimate of the volatility, and Andersen and Bollerslev (1998a) showed that the maximum obtainable $R^2$ from the regression (3), is very small. Hence, there is not necessarily any contradiction between the highly significant parameter estimates and the poor predictive out-of-sample performance, when squared returns are used as measures for the conditional volatility.

To resolve the problem Andersen and Bollerslev (1998a) suggest the use of alternative measures for volatility. Specifically, they show how high frequency data can be used to compute improved ex-post volatility measurements based on cumulative squared intra-day returns. We proceed with this idea, and apply the volatility, estimated from intra-day returns, to evaluate the performance of the volatility models, using the criteria (3)–(11).

### 3.1 Computing Realized Volatility

We adopt a notation similar to the one of Andersen and Bollerslev (1998a). They define the discretely observed series of continuously compounded returns with $m$ observations per day as

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

In this notation $r_{(1),t}$ equals the inter-daily 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 $\sigma_t^2$. This can be seen from the identity

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

so provided that the intra-day returns are uncorrelated we have the identity

$$\sigma_t^2 \equiv \operatorname{var}(r_t|\mathcal{F}_{t-1}) = \sum_{j=1}^{m} \operatorname{var}(r_{(m),t+j/m}|\mathcal{F}_{t-1}). \tag{12}$$

Since $E(r_{(m),t+j/m}|\mathcal{F}_{t-1})$ is typically negligible, we have

$$E(r_{(m),t+j/m}^2|\mathcal{F}_{t-1}) \simeq \operatorname{var}(r_{(m),t+j/m}|\mathcal{F}_{t-1}). \tag{13}$$

Equations (12) and (13) motivate the use of intra-day returns to estimate $\sigma_t^2$. If (13) holds with equality, then an unbiased estimator of $\sigma_t^2$ 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 the $m$-frequency of realized daily volatility.

Several assets are not traded continuously because the market is closed overnight and over weekends. So in several situation, we are only able to observe $f$ of the $m$ possible returns, say the first $f$, given by $r^2_{(m),t+j/m}$, $j = 1, \ldots, f$. In this case we define

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

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

Generally, $E(\hat{\sigma}^2_{(m,f),t}) < E(r^2_t)$ $(= E(\hat{\sigma}^2_{(m),t}))$, so $\hat{\sigma}^2_{(m,f),t}$ is not an unbiased estimator of $\sigma^2_t$. However, if $E(r^2_t)/E(\hat{\sigma}^2_{(m,f),t}) = c$ (does not depend on $t$) then we can use $\hat{c} \cdot \hat{\sigma}^2_{(m,f),t}$ as an estimator of $\sigma^2_t$, where $\hat{c}$ is a consistent estimator of $c$. If intra-day returns are homoskedastic, then $c$ is simply equal to the inverse of the fraction of the day in which we observe intra-day returns, that is $c = m/f$. So if one is willing to make this assumption, then $\hat{c} = m/f$ can be used to scale $\hat{\sigma}^2_{(m,f),t}$.

The use of intra-day returns to estimate the volatility can increase the precision of the estimate of $\sigma^2_t$, dramatically.

**Proposition 1** *Let* $\omega^2 \equiv \text{var}(r^2_t|\mathcal{F}_{t-1})$ *denote the variance of the intra-day estimate of* $\sigma^2_t$, *and suppose that the intra-day returns,* $r_{(m),t+j/m}$, *are independent and Gaussian distributed with mean zero and variance* $\sigma^2_{t+j/m}$, $j = 1, \ldots, m$.

*Then* $\text{var}(\hat{\sigma}^2_{(m),t}) < \omega^2$, *and if the intra-day returns are homoskedastic, i.e.,* $\sigma^2_{t+j/m} = \sigma^2_t/m$, *then* $\text{var}(\hat{\sigma}^2_{(m),t}) = \omega^2/m$. *In particular, the variance of* $\hat{\sigma}^2_{(m,f),t}$ *is only* $1/f$ *times the variance of* $\hat{\sigma}^2_{(1),t}$.

**Proof.** From the identity

$$r^2_t = \sum_{i=1}^{m}\sum_{j=1}^{m} r_{t+i/m} r_{t+j/m},$$

we have that

$$\text{var}(r^2_t|\mathcal{F}_{t-1}) = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} \text{cov}(r_{t+i/m} r_{t+j/m}, r_{t+k/m} r_{t+l/m}|\mathcal{F}_{t-1}).$$

Since the intra-day returns are assumed to be independent with mean zero, then only the terms

that contain pairs of the indices are non-zero. E.g., if $i$ is different from $j$, $k$, and $l$, then

$$\begin{aligned}\operatorname{cov}(r_{t+i/m}r_{t+j/m}, r_{t+k/m}, r_{t+l/m}|\mathcal{F}_{t-1}) &= E(r_{t+i/m}r_{t+j/m}r_{t+k/m}r_{t+l/m}|\mathcal{F}_{t-1}) \\ &= E(r_{t+i/m}|\mathcal{F}_{t-1})E(r_{t+j/m}r_{t+k/m}r_{t+l/m}|\mathcal{F}_{t-1}) \\ &= 0.\end{aligned}$$

The terms that involve two different pairs, contribute

$$E(r^2_{t+i/m}r^2_{t+j/m}|\mathcal{F}_{t-1}) = \sigma^2_{t+i/m}\sigma^2_{t+j/m}, \qquad i \neq j,$$

and the terms that contain the same elements contribute

$$E(r^4_{t+i/m}|\mathcal{F}_{t-1}) = 3\sigma^4_{t+i/m},$$

since $r_{t+i/m}$ is assumed to be Gaussian distributed.

The number of terms that contain two pairs is given by $3m^2$, of which $m$ are the terms with $r^4_{t+i/m}$ (two identical pairs). So the variance estimate of the inter-day estimate of $\sigma^2_t$, is given by

$$\begin{aligned}\operatorname{var}(\hat{\sigma}^2_{(1),t}|\mathcal{F}_{t-1}) &= \sum_{i=1}^{m} 3\sigma^4_{t+i/m} + 3\sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m}\sigma^2_{t+i/m}\sigma^2_{t+j/m}, \\ &= 3\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma^2_{t+i/m}\sigma^2_{t+j/m}.\end{aligned}$$

The variance of the intra-day estimate, $\hat{\sigma}^2_{(m),t} \equiv \sum_{j=1}^{m} r^2_{(m),t+j/m}$, is given by

$$\operatorname{var}(\hat{\sigma}^2_{(m),t}|\mathcal{F}_{t-1}) = \sum_{j=1}^{m}\operatorname{var}(r^2_{(m),t+j/m}|\mathcal{F}_{t-1}) = \sum_{i=1}^{m} 3\sigma^4_{t+i/m}.$$

So using $\hat{\sigma}^2_{(m),t}$ as an estimator of $\sigma^2_t$ rather than $\hat{\sigma}^2_{(1),t} = r^2_t$, reduces the variance by

$$3\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m}\sigma^2_{t+i/m}\sigma^2_{t+j/m},$$

which is generally positive, unless $r_t = r_{t+i/m}$ for some $i$, with probability 1.

Further, if the intra-day returns are homoskedastic, $\sigma^2_{t+i/m} = \sigma^2_{t+j/m}$ for all $i, j = 1, \ldots m$, then it follows that $\sigma^2_{t+i/m} = \sigma^2_t/m$, and the expression for $\operatorname{var}(\hat{\sigma}^2_{(m),t}|\mathcal{F}_{t-1})$ simplifies to

$$\operatorname{var}(\hat{\sigma}^2_{(m),t}|\mathcal{F}_{t-1}) = 3m\left(\frac{\sigma^2_t}{m}\right)^2 = 3\frac{\sigma^4_t}{m},$$

which is only $1/m$ times the variance of $\hat{\sigma}^2_{(1),t}$, which is given by $\operatorname{var}(\hat{\sigma}^2_{(1),t}|\mathcal{F}_{t-1}) = 3\sigma^4_t$.

If only a fraction of the intra-day returns are observed, then the variance of $(m/f) \cdot \hat{\sigma}^2_{(m,f),t}$ is given by

$$\text{var}(\frac{m}{f}\hat{\sigma}^2_{(m,f),t}|\mathcal{F}_{t-1}) = \left(\frac{m}{f}\right)^2 \sum_{i=1}^{f} 3\left(\frac{\sigma_t^2}{m}\right)^2 = 3\frac{\sigma_t^4}{f},$$

which completes the proof. ■

The reduction in the variance of the partial intra-day estimate of $\sigma_t^2$ relies to some extent on the assumption of homoskedasticity. If $\sigma^2_{t+i/m}$ varies with $i$, such that an estimate of $c = E(r_t^2)/E(\hat{\sigma}^2_{(m,f),t})$ is required, then additional variance is added to the partial intra-day estimate of $\sigma_t^2$. In particular, if $f$ is very small and the estimate of $c$ has a large variance, then it can be better to use $r_t^2$ as an estimate of $\sigma_t^2$, rather than creating an estimate from $\hat{\sigma}^2_{(m,f),t}$.

### 3.2 Exchange rate data

Our exchange rate out-of-sample data[5] are identical to the ones used in Andersen and 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 sampled realized daily volatility is computed as $\hat{\sigma}^2_{(288),t}$. This is the measure of volatility that is compared to the models' forecast of volatility, denoted by $h^2_{\cdot,t}$. The significance of relative performance across models is then evaluated using the test for data snooping.

In the technical appendix we list the $R^2$s (denoted $R^2_{\text{inter}}$ and $R^2_{\text{intra}}$) from the regressions corresponding to (3) and (4) for $m = 1, 288$, that is

$$\hat{\sigma}^2_{(1),t} = a + bh^2_{k,t} + u_t \tag{14}$$

$$\hat{\sigma}^2_{(288),t} = a + bh^2_{k,t} + u_t. \tag{15}$$

We find that $R^2_{\text{inter}}$ is typically between 2 and 4 per cent, a very small figure compared to $R^2_{\text{intra}}$, which typically lies between 35 and 45 per cent. We also computed the $R^2$ (denoted $R^{\star 2}_{\text{inter}}$ and $R^{\star 2}_{\text{intra}}$) 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 was maintained. The estimated intra-day volatilities, used in the comparison, are given by $\hat{\sigma}_t^2 = .8418\,\hat{\sigma}^2_{(288),t}$.

[5] This data set was kindly provided by Tim Bollerslev. For the construction of the series and additional information, we refer to Andersen and Bollerslev (1997, 1998b) and Andersen, Bollerslev, Diebold, and Labys (2000)

The reason for the scaling is explained in the next subsection. Intra-day volatility and returns are plotted in Figure 2.

### 3.3 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 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, and Sosebee (1993) document NYSE trading and quoting procedures. In this application we only consider IBM stock prices. 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 the market-microstructure of the Stock Exchange into account. Factors, such as the bid-ask spreads and the irregular spacing of price quotes, could potentially distort our estimates of volatility, if such estimates were based on tick-by-tick data. Andersen and Bollerslev (1997, 1998a, 1998b) and Andersen, Bollerslev, Diebold, and Ebens (2000) circumvented this obstacle by estimating the volatility from artificially constructed five-minutes returns. We take a similar approach, in the sense that we fit a cubic spline through all daily 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[6]. A random sample of these splines, as well as mid quotes, are displayed in Figure 1. From the splines we extract artificial one- and five-minute returns, which leads to a total of $f_1 = 390$ one-minute returns or $f_5 = 78$ five-minute returns for each of the days. This delivers our measure of realized volatility. Because we only have 390 one-minute returns of the $m_1 = 1,440$ theoretical one-minute returns, and similarly we only have 78 of the 288 theoretical five-minute returns, we denote our measure for the volatility by

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

where $(m, f) = (1440, 390)$ for the one-minute returns and $(m, f) = (288, 78)$ for the five-minute returns.

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

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

The intra-day measures, $\hat{\sigma}^2_{(1440,390),t}$ and $\hat{\sigma}^2_{(288,78),t}$, are not directly comparable to the inter-day measure, $\hat{\sigma}^2_{(1),t}$, because they are calculated from a proportion of the 24 hours in a day. So, we need to adjust for this bias in order to avoid a distortion of the evaluation based on the loss functions (5)–(11).

It is interesting to note that this bias will 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 the evaluation. 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 the model with the highest $R^2$, that is better than any affine transformation of the model with the smallest $R^2$. Since the "optimal" affine transformation is only known ex-post, it is not necessarily a good criterion for comparison of volatility models.

Thus, in order to make the loss function relevant for the comparison, we need to adjust for the mismatch between the volatility estimated from (a fraction of) the intra-day returns, and the inter-day returns. 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. We therefore prefer to re-scale our intra-day estimated measure for volatility. It seems natural to scale $\hat{\sigma}^2_{(m,f),t}$ by a number that is inversely proportional to the fraction of the day we extract data from, i.e., a scaling by $\frac{f}{m}$. However, it is not obvious that an hour in which the market is open should be weighted equally to an hour in which the market is closed. Therefore we choose to scale $\hat{\sigma}^2_{(m,f),t}$ such that its sample average equals the sample average of $\hat{\sigma}^2_{(1),t}$.

Thus, we define

$$\hat{\sigma}^2_t \equiv \hat{c} \cdot \hat{\sigma}^2_{(m,f),t},$$

where

$$\hat{c} = \left( \frac{\sum_{t=1}^{n} \hat{\sigma}^2_{(1),t}}{\sum_{t=1}^{n} \hat{\sigma}^2_{(m,f),t}} \right), \tag{16}$$

as our measure for the volatility on day $t$, $t = 1, \ldots, n$.

Although this adjustment is only known ex-post it should not distort our comparison of the models, because the ex-post information is only used in the evaluation and is not included in the information set, which the volatility models apply for their forecast. If, for some reason, there is a difference between $E(\hat{\sigma}^2_{(m,m),t}|\mathcal{F}_{t-1})$ 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) $r_t^2 = \hat{\sigma}^2_{(1),t}$. Thus, a potential difference between $E(\hat{\sigma}^2_{(m,m),t}|\mathcal{F}_{t-1})$ and $E(r_t^2|\mathcal{F}_{t-1})$ is a justification for making an adjustment, of the intra-day estimate of the volatility.

The volatility estimates based on the five-minute returns need to be adjusted by about 4.5, (the value of $\sum_{t=1}^{n} \hat{\sigma}^2_{(1),t} / \sum_{t=1}^{n} \hat{\sigma}^2_{(m,f),t}$) which is a higher correction than $\frac{1440}{78} \approx 3.7$. Thus, the squared five-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 fact that we need to adjust the volatilities by a number different than 3.7 can have several possible explanations. First of all, it could be the result of sample error. However, $n$ is too large in our application for sampling error alone to explain the difference. A second explanation is that autocorrelation in the intra-day returns can cause a bias. This can be seen from the relation

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

If we ignore that only a fraction of the intra-day returns are observed, we have evidence that $\sum_{t=1}^{n} r_t^2 > \sum_{t=1}^{n} \left( \sum_{j=1}^{m} r^2_{t+j/m} \right)$, 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. Such a "positive average correlation" can arise from the market micro-structure, but need not be a real phenomenon, as it could be an artifact of the way we created 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 could be that returns are relatively more volatile between close and open, than between open and close, measured per unit of time. This explanation is plausible

if relatively more information arrived to the market while it is closed. Market micro-structures that leave fewer opportunities to hedge against risk while the market is closed, may also cause a higher volatility while the market is closed. However, this explanation requires the additional presumption that hedging against risk has a stabilizing effect on the market.

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 $E(r_{t+i/m}|\mathcal{F}_{t-1})$, $i = 1, \ldots, m$. Suppose that $E(r_{t+i/m}|\mathcal{F}_{t-1}) = 0$ for $i = 1, \ldots, f$, but is positive during the time the market is closed. Then $r_t^2$ would, on average, be larger than $\frac{m}{f}\sum_{i=1}^{f} r_{t+i/m}^2$, even if 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 causes the difference, as long as our adjustment does not favor some models over others. Since the adjustment is made ex-post and independent of the forecasts of the models, the adjustment should not matter for our comparison. The adjustment of the partial intra-day estimated volatilities, is $\hat{\sigma}_t^2 = 4.4938\,\hat{\sigma}^2_{(288,78),t}$, where $\hat{c} = 4.4938$ is calculated using (16). This is the measure we apply in the evaluation, and the estimated intra-day volatilities are plotted in Figure 3 along with the daily returns.

# 4 The Bootstrap Implementation

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

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

The parameters of the volatility models are estimated using the first $R$ observations, and these parameter estimates are then used to make the forecasts for the remaining $n$ observations. Let $l + 1$ denote the number of competing forecasting models. The $k$'th model yields the forecasts

$$h_{k,1}^2, \ldots, h_{k,n}^2, \qquad k = 0, 1, \ldots, l,$$

that are compared to the intra-day calculated volatility

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

The forecast $h^2_{\cdot,t}$ of the realized volatility $\sigma^2_t$ leads to the utility $u(\hat{\sigma}^2_t, h^2_{\cdot,t})$, where $u$ is defined from the performance measures listed in Section 3, e.g., $u(\hat{\sigma}^2_t, h^2_{\cdot,t}) = -(\hat{\sigma}^2_t - h^2_{\cdot,t})^2$ for the mean squared forecast error criterion.

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

$$X_{kt} \equiv u_{k,t} - u_{0,t}, \qquad k = 1, \ldots, l, \quad t = 1, \ldots, n.$$

The expected performance of model $k$ relative to the benchmark is defined as

$$\lambda_k \equiv E\left[X_{kt}\right], \qquad k = 1, \ldots, l.$$

Note that this parameter is well-defined (independent of $t$) due to the assumed stationarity of $\hat{\sigma}^2_t$ and $h^2_{\cdot,t}$.

A model that outperforms the benchmark model, model $k^*$ say, translates into a positive value of $\lambda_{k^*}$. Thus, we can analyze whether any of the competing models significantly outperform the benchmark model, by testing the null hypothesis that $\lambda_k \leq 0$, $k = 1, \ldots, l$. That is that none of the models are better than the benchmark. If we reject this hypothesis, we have evidence for the existence of a model that is better than the benchmark model. We can reformulate the null hypothesis to the equivalent hypothesis:

$$H_0 \qquad \lambda_{\max} \equiv \max_{k=1,\ldots,l} \lambda_k \leq 0.$$

We can, by the law of large numbers, estimate the parameter, $\lambda_k$, with the sample average $\bar{X}_{n,k} = n^{-1} \sum_{t=1}^{n} X_{kt}$, and $\lambda_{\max}$ is therefore consistently estimated by $\bar{X}_{n,\max} \equiv \max_{k=1,\ldots,l} \bar{X}_{n,k}$, which measures how well the best model performed compared to the benchmark model. Even if $\lambda_{\max} \leq 0$ it can (and will) by chance happen that $\bar{X}_{n,\max} > 0$. The relevant question is whether $\bar{X}_{n,\max}$ is too large for it to be plausible that $\lambda_{\max}$ is truly non-positive. This is precisely what the test for data snooping is designed to answer. The test for data snooping estimates the distribution of $\bar{X}_{n,\max}$ under the null hypothesis, and from this distribution we are able to evaluate whether $\bar{X}_{n,\max}$ 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 that is significantly better that the benchmark.

We can describe the performance of the $l$ models relative to the benchmark by the $l$-dimensional vector $\mathbf{X}_t = \left(X_{1t}, \ldots, X_{l,t}\right)'$, $t = 1, \ldots, n$, and the sample performance is given

by $\overline{\mathbf{X}}_n = n^{-1} \sum_{t=1}^{n} \mathbf{X}_t$. The fundamental assumption that enables the test for data snooping to test the significance, is that $\bar{X}_{n,\max}$ (appropriately scaled) converges in distribution. If $\{X_t\}$ satisfies assumptions such that a central limit theorem applies, we have that

$$n^{1/2}(\overline{\mathbf{X}}_n - \lambda) \xrightarrow{d} N_l(\mathbf{0}, \mathbf{\Omega}), \tag{17}$$

where "$\xrightarrow{d}$" denotes convergence in distribution and where $\lambda = (\lambda_1, \ldots, \lambda_l)'$ and

$$\mathbf{\Omega} \equiv E\left[(\mathbf{X}_t - \lambda)(\mathbf{X}_t - \lambda)'\right].$$

So as $n \to \infty$, $\overline{\mathbf{X}}_n$ is "close" to $\lambda$, and by Slutsky's theorem, it holds that $\bar{X}_{n,\max} \equiv \max_k \bar{X}_{n,k}$ is "close" to $\lambda_{\max}$. Therefore, a large positive value of $\bar{X}_{\max}$ indicates that the benchmark model is outperformed. The tests for data snooping (tests for superior predictive ability) of White (2000) and Hansen (2001) applies the result in (17) to derive a critical value for $\bar{X}_{\max}$, and this critical value is the threshold at which $\bar{X}_{\max}$ becomes too large for it to be plausible that $\lambda_{\max} \leq 0$.

### 4.1 Bootstrap Implementation

The bootstrap implementation of the tests for data snooping is constructed such that it generates $B$ draws from the distribution $N(\lambda, \mathbf{\Omega})$, where $\lambda$ satisfies the null hypothesis, i.e., $\lambda \leq \mathbf{0}$. These draws are used to approximate the distribution of $\bar{X}_{\max}$, from which critical values and $p$-values are derived.

First, let $b = 1, \ldots, B$ index the re-samples of $\{1, \ldots, n\}$, given by $\theta_b(t)$, $t = 1, \ldots, n$. 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 and Buchinsky (2000). We apply the stationary bootstrap of Politis and Romano (1994), where $\theta_b(t)$ is constructed by combining blocks with random length that are geometrically distributed with parameter $q \in (0, 1]$. The parameter $q$, is used to preserve possible time-dependence in $X_k(t)$. The re-samples are generated as follows:

1. Initiate the random variable, $\theta_b(0)$, as uniform distribution on $\{1, \ldots, n\}$.
2. For $t = 1, \ldots, n$

   Generate $u$ uniformly on $[0, 1]$.

(a) If $u$ is smaller than $q$, then the next observation is chosen uniformly on $\{1, \ldots, n\}$, 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:

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

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

$$X^*_{k,b}(t) \equiv X_k(\theta_b(t)) - g(\bar{X}_{n,k}), \qquad b = 1, \ldots, B,\ t = 1, \ldots, n,$$

where $g$ is a data dependent constant, specified below, that ensures that $X^*_{k,b}(t)$ satisfy the null hypothesis, i.e., $E(X^*_{k,b}(t)) \leq 0$.

The sample average of the resamples are given by

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

and by the construction of the bootstrap re-samples, the sample averages satisfies the null hypothesis, $E\left(\bar{X}^*_{n,k,b} | \mathbf{X}_1, \ldots, \mathbf{X}_n\right) \leq 0$, and consequently, $E\left(\bar{X}^*_{n,\max,b} | \mathbf{X}_1, \ldots, \mathbf{X}_n\right) \leq 0$, where

$$\bar{X}^*_{n,\max,b} \equiv \max_{k=1,\ldots,l} \bar{X}^*_{n,k,b},\ b = 1, \ldots, B.$$

The property of the bootstrap, is that each of the resamples, $\bar{X}^*_{n,\max,b}$, approximates a random draw of $\bar{X}_{n,\max}$, according to a probability law that is consistent with the null hypothesis. If the null hypothesis is true and if the bootstrap-generated variables $\bar{X}^*_{n,\max,b}$ are like random draws of $\bar{X}_{n,\max}$, then $\bar{X}_{n,\max}$ would rarely be extreme (very large) relative to the values $B$ draws, given by $\bar{X}^*_{n,\max,b}$, $b = 1, \ldots, B$.

Different choices of $g$ will lead to a different bootstrap distribution, which is consistent with the null hypothesis. The Reality Check of White (2000), $\mathrm{DS}_u$, applies $g(\bar{X}_{n,k}) = \bar{X}_{n,k}$, whereas the $\mathrm{DS}_l$ and $\mathrm{DS}_c$, of Hansen (2001), applies $g(\bar{X}_{n,k}) = \max(\bar{X}_{n,k}, 0)$ and

$$g(x) = g(x, A_{n,k}) = \begin{cases} 0 & \text{if } x \leq -A_{n,k} \\ x & \text{otherwise,} \end{cases}$$

respectively, where $A_{n,k}$ is given below.

The choice of $g$ made by the $\text{DS}_u$ corresponds to an assumption that $\lambda$ equals $\mathbf{0}$ under the null hypothesis. This makes the test conservative and its $p$-value can be viewed as an upper bound for the true $p$-value. The $\text{DS}_l$ is a liberal test that provides a lower bound for the $p$-value, and the $\text{DS}_c$ provides a consistent $p$-value[7]. The consistency of the $\text{DS}_c$s $p$-value is achieved by the correction factor, $A_{n,k}$, which must be constructed such that it vanishes asymptotically, $A_{n,k} \xrightarrow{p} 0$. However, the rate at which it vanishes must be slow enough such that, as $n \to \infty$, we are able to determine exactly the models for which $\mu_k = 0$. This is important to obtain the consistency, because the models with $\mu_k < 0$ do not have an influence on the distribution of $\bar{X}_{n,\max}$ in the limit. So even though both the $\text{DS}_l$ and the $\text{DS}_c$ apply consistent estimators for $\lambda_k$ under the null hypothesis[8], only the $\text{DS}_c$ achieves generally consistent $p$-values. The $p$-values of the Reality Check, $\text{DS}_u$, are generally inconsistent. Only if $\mu_k = 0$, for all $k = 1, \ldots, l$, are the $p$-values of the $\text{DS}_u$ consistent.

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

$$A_{n,k} \equiv \frac{1}{4} n^{1/4} \sqrt{\widehat{\text{var}}(\bar{X}_{n,k})}, \tag{18}$$

which requires an estimate of $\text{var}(\bar{X}_{n,k})$. Simpler choices are available, for example $A_{n,k} = n^{-1/4}$ 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. The expression in (18) is straightforward to implement, since the variance estimate is easily obtained from the bootstrap re-samples

$$\widehat{\text{var}}(\bar{X}_{n,k}) = B^{-1} \sum_{b=1}^{B} \left[ \left( n^{-1} \sum_{t=1}^{n} X_k(\theta_b(t)) \right) - \bar{X}_{n,k} \right]^2,$$

where $\bar{X}^*_{n,k,b} = n^{-1} \sum_{t=1}^{n} X^*_k(\theta_b(t))$.

From the bootstrap generated draws of $\bar{X}_{n,\max}$, given by $\bar{X}^*_{n,\max,1}, \ldots, \bar{X}^*_{n,\max,B}$, we can evaluate whether $\bar{X}_{n,\max}$ is an extreme observation or not. If we conclude that the observation of $\bar{X}_{n,\max}$ is extreme, (is too large), then we have evidence against the null hypothesis, and can conclude that an alternative model has a superior predictive ability, compared to that of the benchmark model.

[7] The subscripts, $u$, $c$, and $l$, refer to "upper bound", "consistent", and "lower bound".

[8] The $\text{DS}_l$ applies $\hat{\lambda}_k = \min(\bar{X}_{n,k}, 0)$ as an estimator for $\mu_k$ and the $\text{DS}_c$ applies $\hat{\lambda}_k = \bar{X}_{n,k} 1(\bar{X}_{n,k} < A_{n,k})$, where $1(\cdot)$ is the indicator function.

The data snooping $p$-value, $p_{ds}$ is given by

$$p_{ds} \equiv \sum_{b=1}^{B} \frac{1(\bar{X}^*_{n,\max,b} > \bar{X}_{n,\max})}{B},$$

where $1(\cdot)$ is the indicator function. So if relatively few, or none, of the bootstrap draws $\bar{X}^*_{n,\max,b}$ are larger than the observed value, then $\bar{X}_{n,\max}$ is an extreme observation, and has a low $p$-value. Thus a low $p$-value corresponds to a situation where the best alternative model is so much better than the benchmark, that it is unlikely to be a result of luck.

This procedure is repeated for each of the three tests for data snooping, 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 consistent test for data snooping, $\mathrm{DS}_c$, will depend on the actual choice of correction factors $A_{n,k}$, $k = 1, \ldots, l$. 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, $A_{n,k}$.

## 5 Results from the Analysis

The models were compared using two different benchmark models. The two benchmark models in our analysis are the ARCH(1) and GARCH(1,1) models. Our results are given in Tables 3 and 4.

When the ARCH(1) model is chosen as the benchmark model, it is clearly outperformed by alternative models. Once we choose the GARCH(1,1) model as the benchmark, the $p$-values of tests for data snooping increases dramatically, due to the better performance by the GARCH(1,1). For the exchange rate data the GARCH(1,1) seems to be able to capture the variation in the conditional variance. Its performance is not statistically worse than any of the competing models. For the IBM data the answer is less obvious. One $p$-value is as low as .04, and several are about .10. So statistically there is some evidence that a better forecasting model exists.

It is interesting to see how the $p$-values of the three tests for data snooping differ in some cases. When we analyze the data using the ARCH(1) model as the benchmark, the $p$-values mostly agree. But in the case where the GARCH(1,1) model is the benchmark model, the $p$-

values differ quite substantially. The reason is that the $DS_u$ of White (2000) is sensitive to inclusion of poor models, see Hansen (2001). When we use the GARCH(1,1) and the benchmark model, there are several models that are considerably worse performing relative to the GARCH(1,1). This hurts the $DS_u$, and its $p$-values are no longer consistent for the true $p$-values. The $p$-values of the $DS_c$ remain consistent (under the null hypothesis).

It is worth mentioning that the power properties of the tests for data snooping can be poor, is some situations. So the fact that we fail to find a model that is significantly better than the GARCH(1,1) may be explained by this lack of power. In other words, the sample size, $n$, of our out-of-sample data may be too short for the tests for data snooping to significantly detect that a better model exists. Additional information may be obtained from the relative ranking of the models, which are listed in Tables 5–10. 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 are given the scores 100, 0, and 50 respectively. Since we use 7 criteria for each of the two data sets, each model has 14 scores. The last column in the tables is the average of the 14 scores.

As can be seen from the Tables 5–10, the ARCH(1) model is generally amongst the worst models. This is true for every of the six models that uses the ARCH(1) specification for the volatility process. However, in the analysis of the IBM data, there are about 25% of the volatility models that performs 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 not quite as good in the IBM data. It is interesting to notice that it is not the same models that do well in the two data sets, not 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, although most of the better models have GARCH-in-mean. The overall best GARCH(2,2) model is the one with $t$-distributed errors and GARCH-in-mean, see Table 10, the overall best LOG-GARCH(2,2) model is the model with Gaussian errors and either zero-mean or a GARCH-in-mean, see Tables 5 and 7, and the best GQ-ARCH(2,1) model is the model with Gaussian errors and GARCH-in-mean, see Table 7.

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 with $t$-distributed errors and mean zero, see Table 8. Also the V-GARCH specification does quite well, in particular in terms of the two MAD criteria, that 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 $(p, q) = (1, 2)$. Note 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 evaluations.

The fact that the EGARCH specification performs far better using $t$-distributed standardized errors, rather than Gaussian, shows the importance of modelling the entire distribution. It is not sufficient to focus on the specification of the volatility, although it (in our analysis) is the only object of interest.

The IGARCH specifications are surprisingly poor, for all but the $PSE$, $(L_3)$, criterion. In terms of this criterion the model does quite well. The difference of the relative performance (across criteria) is most likely due to events where the IGARCH predicted a very large volatility. A large misprediction, ($h^2_{k,t}$ too large) would result in a large value of most loss functions. However, the loss of over-predicting the volatility cannot exceed *one* when the PSE is applied, thus over-predictions have a small weight relative to under-predictions when this loss function is applied. The PSE loss function, as defined by Bollerslev, Engle, and Nelson (1994), measures percentage squared error relative to the predicted volatility[9], $h^2_{k,t}$. It may be this property that helps the IGARCH in terms of its relative performance when the PSE is applied.

Similarly, the NGARCH(2,2) with Gaussian errors and a zero mean specification is the best model in terms of the PSE criterion, but in the bottom 10% with respect to the outlier-robust $\text{MAD}_i$ criteria, $i = 1, 2$, (in the analysis of the IBM data). The opposite is the case for some of the V-GARCH models.

The fact that the relative performance varies substantially with the choice of loss function emphasizes how important it is to use the appropriate loss function, in applied work. However, based on our observation with respect to the percentage squared error, it seems more reasonable to measure percentage errors relative to the intra-day estimated measure of $\sigma^2_t$, whenever such

[9]To measure mispredictions relative to the prediction itself seems rather awkward. However, unless intra-day returns are used, $h^2_t$ is typically the best estimate of $\sigma^2_t$ and far better than using the noisy squared returns, $r^2_t$.

an estimate is available. Hence, we argue that $\widetilde{PSE} = n^{-1}\sum_{t=1}^{n}\left(\hat{\sigma}_t^2 - h_t^2\right)^2 \hat{\sigma}_t^{-4}$ is a more appropriate loss function, than (7).

# 6 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 predictive ability, where the forecasts of the different models are compared to intra-day estimated measures of realized volatility. The intra-day estimated volatilities provide good estimates of realized volatility, which makes the comparison of different volatility models more precise.

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 data snooping, $\mathrm{DS}_c$, of Hansen (2001).

If we compare the estimated volatility models to a simple ARCH(1) model, we find the ARCH(1) to be significantly outperformed by other models. That is, there is strong evidence that significant gains in forecasting ability can be obtained by using a competing model. This does not come as a surprise to those familiar with volatility models, because the ARCH(1) model is not flexible enough to capture the persistence in volatility. In contrast to the ARCH(1), we do not find much evidence that the GARCH(1,1) model is outperformed. When the family of competing models are compared to the GARCH(1,1) model, we cannot reject that none of the competing models are better than the GARCH(1,1). This is somewhat surprising, because the GARCH(1,1) model corresponds to a simple news impact curve, and a GARCH(1,1) process cannot generate a leverage effect.

However, it may be that our lack of strong evidence against the GARCH(1,1) model can be explained by the limitations of our analysis. First, it may be that a comparison using other assets would result in a different conclusion. For example, one or more of the competing models may significantly outperform the GARCH(1,1), if the models are compared using returns of stock indices or bonds. Secondly, there might be a model, not included in our analysis, which is indeed better than the GARCH(1,1). Although we estimated 330 different models we have not entirely exhausted the space of volatility models. For example, we could add models that combine the forecast of two or more volatility models. Thirdly, the power of the test for data snooping can, in some situations, be poor. If this is relevant to our applications, then a longer

sample could result in a significant outperformance of the benchmark model. However, the test for data snooping, $DS_c$, is not powerless in our analysis. This is shown by the fact that the $DS_c$ finds the ARCH(1) model to be significantly outperformed.

Our subsequent analysis leads to some interesting ideas. It seems plausible that volatility models are good at predicting the intra-day volatility. This is an accomplishment in itself, because they are estimated using a much smaller information set, that primarily contains inter-day returns. Therefore it would be interesting to analyze if better forecasts can be constructed from models that are not limited to using inter-day returns. In particular models that apply an intra-day estimated measure of volatility may provide more accurate forecasts of volatility. Or more generally, models that include information provided by intra-day returns may provide superior forecasts of the distribution of $r_t$. We leave this for future research.

## References

ANDERSEN, T. G., AND T. BOLLERSLEV (1997): "Intraday periodicity and volatility persistence in financial markets," *Journal of Empirical Finance*, 4, 115–158.

——— (1998a): "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts," *International Economic Review*, 39(4), 885–905.

——— (1998b): "Deutsche mark-dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies," *Journal of Finance*, 53(1), 219–265.

ANDERSEN, T. G., T. BOLLERSLEV, F. X. DIEBOLD, AND H. EBENS (2000): "The distribution of stock return volatility," *Forthcomming Journal of Financial Economics*.

ANDERSEN, T. G., T. BOLLERSLEV, F. X. DIEBOLD, AND P. LABYS (2000): "The distribution of exchange rate volatility," *Forthcoming Journal of the American Statistical Association*.

ANDREWS, D. W. K., AND M. BUCHINSKY (2000): "A Three-Steep Method for Choosing the Number of Bootstrap Repetitions," *Econometrica*, 68, 23–52.

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.

——— (1987): "A conditional heteroskedastic time series model for speculative prices and rates of return," *Review of Economics & Statistics*, 69(3), 542–547.

BOLLERSLEV, T., R. F. ENGLE, AND D. NELSON (1994): "ARCH models," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. IV, pp. 2961–3038. Elsevier Science B.V.

DIEBOLD, F. X., AND J. A. LOPEZ (1996): "Forecast Evaluation and Combination," in *Handbook of Statistics*, ed. by G. S. Maddala, and C. R. Rao, vol. 14: Statistical Methods in Finance, pp. 241–268. North-Holland, Amsterdam.

DING, Z., C. W. J. GRANGER, AND R. F. ENGLE (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., D. V. LILLEN, AND R. P. ROBINS (1987): "Estimating time varying risk premia in the term structure: The ARCH-M model," *Econometrica*, 55, 391–407.

ENGLE, R. F., AND V. NG (1993): "Measuring and testing the impact of news on volatility," *Journal of Finance*, 48, 1747–1778.

ENGLE, R. F., AND A. J. PATTON (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.

GALLANT, A. R., AND G. TAUCHEN (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 Review*, 5, 57–61.

GLOSTEN, L. R., R. JAGANNATHAN, AND D. E. RUNKLE (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., AND B. W. SILVERMAN (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," *http://chico.pstc.brown.edu/~phansen*.

HARVEY, C. R., AND A. SIDDIQUE (1999): "Autoregressive conditional skewness," *Journal of Financial and Quantitative Analysis*, 34(4), 465–487.

HASBROUCK, J., G. SOFIANOS, AND D. SOSEBEE (1993): "Orders, Trades, Reports and Quotes at the New York Stock Exchange," Discussion paper, NYSE, Research and Planning Section.

HENTSHEL, L. (1995): "All in the family: Nesting symmetric and asymmetric GARCH models," *Journal of Financial Economics*, 39, 71–104.

HIGGINS, M. L., AND A. K. BERA (1992): "A class of nonlinear ARCH models," *International Economic Review*, 33, 137–158.

PAGAN, A. R., AND G. W. SCHWERT (1990): "Alternative models for conditional volatility," *Journal of Econometrics*, 45, 267–290.

PANTULA, S. G. (1986): "Modelling persistence in conditional variances: A comment," *Econometric Review*, 5, 71–74.

POLITIS, D. N., AND J. P. ROMANO (1994): "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

PRESS, W. H., S. A. TEUKOLSKY, W. T. VETTERLING, AND B. P. FLANNARY (1992): *Numerical Recipes in C.* : Cambrigde University Press2 edn.

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.

——— (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.

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 $(\sigma^2)$ | $\mu_t = \mu_0 + \mu_1 \sigma^2_{t-1}$ |

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

[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)

| | Benchmark: ARCH(1) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Criterion | Performance | | | | $p$-values | | | |
| | Bench. | Worst | Median | Best | Naive | $DS_l$ | $DS_c$ | $DS_u$ |
| $MSE_2$ | −.1288 | −.1404 | −.0853 | −.0778 | .0420 | .0955 | .0990 | .0990 |
| $MSE_1$ | −.0463 | −.0492 | −.0339 | −.0314 | .0085 | .0270 | .0295 | .0295 |
| PSE | −.3725 | −.4583 | −.2052 | −.1868 | .0635 | .1140 | .1685 | .1685 |
| QLIKE | −.3747 | −.3795 | −.3332 | −.3252 | .0080 | .0200 | .0200 | .0200 |
| R2LOG | −.4124 | −.4250 | −.3366 | −.3154 | .0005 | .0035 | .0045 | .0045 |
| $MAD_2$ | −.2533 | −.2904 | −.2194 | −.2045 | .0010 | .0075 | .0150 | .0160 |
| $MAD_1$ | −.1698 | −.1834 | −.1473 | −.1396 | .0000 | .0045 | .0045 | .0050 |

| | Benchmark: GARCH(1,1) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Criterion | Performance | | | | $p$-values | | | |
| | Bench. | Worst | Median | Best | Naive | $DS_l$ | $DS_c$ | $DS_u$ |
| $MSE_2$ | −.0812 | −.1404 | −.0853 | −.0778 | .1975 | .5525 | .8330 | .9690 |
| $MSE_1$ | −.0321 | −.0492 | −.0339 | −.0314 | .2870 | .6085 | .7300 | .9835 |
| PSE | −.2010 | −.4583 | −.2052 | −.1868 | .0630 | .3260 | .5285 | .8975 |
| QLIKE | −.3280 | −.3795 | −.3332 | −.3252 | .2655 | .4570 | .5965 | .9755 |
| R2LOG | −.3218 | −.4250 | −.3366 | −.3154 | .0760 | .5430 | .6325 | .9670 |
| $MAD_2$ | −.2107 | −.2904 | −.2194 | −.2045 | .1695 | .4420 | .5720 | .9165 |
| $MAD_1$ | −.1415 | −.1834 | −.1473 | −.1396 | .0645 | .6395 | .7200 | .9855 |

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 DS $p$-values controls for the full model space. The $DS_l$ and $DS_u$ provide a lower and upper bound for the true $p$-values respectively, whereas the $DS_c$ $p$-values are consistent for the true $p$-values.

Table 4: IBM Data

Benchmark: ARCH(1)

| Criterion | Performance | | | | $p$-values | | | |
|---|---|---|---|---|---|---|---|---|
| | Bench. | Worst | Median | Best | Naive | $DS_l$ | $DS_c$ | $DS_u$ |
| $MSE_2$ | −30.9296 | −31.0289 | −24.9773 | −22.1609 | .0065 | .0225 | .0225 | .0225 |
| $MSE_1$ | −0.8047 | −0.8108 | −0.6222 | −0.5599 | .0045 | .0155 | .0155 | .0155 |
| PSE | −2.2086 | −2.2592 | −0.6875 | −0.4607 | .0055 | .0065 | .0065 | .0065 |
| QLIKE | −2.9177 | −2.9237 | −2.7670 | −2.7423 | .0000 | .0005 | .0005 | .0005 |
| R2LOG | −0.4837 | −0.5357 | −0.4016 | −0.3776 | .0115 | .0650 | .0770 | .0770 |
| $MAD_2$ | −3.0774 | −3.5636 | −2.9850 | −2.8111 | .0030 | .1275 | .1760 | .2015 |
| $MAD_1$ | −0.6191 | −0.7092 | −0.5915 | −0.5552 | .0050 | .1010 | .1330 | .1455 |

Benchmark: GARCH(1,1)

| Criterion | Performance | | | | $p$-values | | | |
|---|---|---|---|---|---|---|---|---|
| | Bench. | Worst | Median | Best | Naive | $DS_l$ | $DS_c$ | $DS_u$ |
| $MSE_2$ | −25.2323 | −31.0289 | −24.9773 | −22.1609 | .0435 | .0970 | .0975 | .1415 |
| $MSE_1$ | −0.6317 | −0.8108 | −0.6222 | −0.5599 | .0325 | .1060 | .1585 | .3010 |
| PSE | −0.7474 | −2.2592 | −0.6875 | −0.4607 | .0180 | .0335 | .0405 | .3655 |
| QLIKE | −2.7711 | −2.9237 | −2.7670 | −2.7423 | .0235 | .0980 | .1230 | .3865 |
| R2LOG | −0.4086 | −0.5357 | −0.4016 | −0.3776 | .0170 | .2985 | .3560 | .6365 |
| $MAD_2$ | −3.0307 | −3.5636 | −2.9850 | −2.8111 | .0050 | .0655 | .1175 | .1850 |
| $MAD_1$ | −0.6018 | −0.7092 | −0.5915 | −0.5552 | .0045 | .0480 | .1150 | .1645 |

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 DS $p$-values controls for the full model space. The $DS_l$ and $DS_u$ provide a lower and upper bound for the true $p$-values respectively, whereas the $DS_c$ $p$-values are consistent for the true $p$-values.

Table 5: Models with Gaussian error distribution and mean zero

| Model | Exchange Rate Data | | | | | | | IBM Data | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | |
| ARCH(1) | 4.6 | 4.0 | 1.2 | .9 | 4.3 | 8.2 | 5.5 | 1.5 | 1.5 | 1.5 | 1.5 | 8.2 | 28.6 | 20.7 | 6.6 |
| GARCH(1,1) | 86.0 | 93.6 | 67.8 | 97.6 | 93.3 | 89.7 | 90.9 | 42.6 | 40.7 | 41.3 | 42.2 | 36.8 | 35.0 | 35.3 | 63.8 |
| GARCH(2,1) | 84.5 | 87.5 | 63.2 | 94.8 | 86.6 | 86.6 | 89.1 | 25.5 | 15.2 | 45.9 | 28.6 | 12.5 | 14.3 | 12.8 | 53.4 |
| GARCH(1,2) | 85.1 | 80.5 | 19.8 | 65.3 | 87.5 | 91.8 | 91.8 | 40.7 | 30.7 | 44.7 | 41.6 | 25.2 | 27.7 | 28.9 | 54.4 |
| GARCH(2,2) | 89.1 | 88.4 | 18.2 | 72.6 | 92.1 | 96.4 | 96.4 | 43.2 | 29.2 | 49.2 | 42.6 | 18.2 | 23.7 | 21.6 | 55.8 |
| IGARCH(1,1) | 7.3 | 7.9 | 56.8 | 17.9 | 8.8 | 6.1 | 8.2 | 13.1 | 3.6 | 80.2 | 14.6 | 1.8 | 2.4 | 2.1 | 16.5 |
| IGARCH(2,1) | 6.7 | 7.6 | 50.2 | 17.0 | 8.5 | 5.8 | 7.9 | 15.8 | 7.0 | 71.1 | 17.0 | 5.2 | 5.5 | 5.5 | 16.5 |
| IGARCH(1,2) | 4.0 | 6.1 | 32.2 | 14.6 | 7.9 | 4.6 | 6.4 | 13.7 | 4.3 | 77.8 | 14.9 | 2.1 | 3.0 | 2.7 | 13.9 |
| IGARCH(2,2) | 10.6 | 8.8 | 45.0 | 14.0 | 6.1 | 7.9 | 8.8 | 37.1 | 8.8 | 80.5 | 31.9 | 7.3 | 7.0 | 6.4 | 20.0 |
| TS-GARCH(1,1) | 54.4 | 58.7 | 95.7 | 73.6 | 61.1 | 35.3 | 40.1 | 86.3 | 68.7 | 93.9 | 84.5 | 29.5 | 24.0 | 24.3 | 59.3 |
| TS-GARCH(2,1) | 57.1 | 57.4 | 97.9 | 72.3 | 58.4 | 32.5 | 38.3 | 72.6 | 68.1 | 79.9 | 82.7 | 35.3 | 31.3 | 31.6 | 58.3 |
| TS-GARCH(1,2) | 91.8 | 88.8 | 68.7 | 86.6 | 84.2 | 76.6 | 72.6 | 87.2 | 71.1 | 92.4 | 90.3 | 33.1 | 26.7 | 27.1 | 71.2 |
| TS-GARCH(2,2) | 94.8 | 95.7 | 60.2 | 91.2 | 93.0 | 82.4 | 79.9 | 79.3 | 69.9 | 83.6 | 85.7 | 35.9 | 31.6 | 31.0 | 72.4 |
| A-GARCH(1,1) | 71.7 | 78.1 | 49.5 | 86.0 | 79.6 | 79.6 | 82.7 | 47.1 | 65.0 | 38.6 | 49.5 | 81.5 | 70.5 | 72.3 | 68.0 |
| A-GARCH(2,1) | 60.5 | 59.9 | 29.2 | 54.1 | 62.3 | 67.5 | 74.8 | 36.5 | 38.6 | 29.5 | 30.7 | 66.6 | 58.4 | 62.3 | 52.2 |
| A-GARCH(1,2) | 85.7 | 81.2 | 21.3 | 66.3 | 86.9 | 92.7 | 92.7 | 45.0 | 63.8 | 34.7 | 46.5 | 86.0 | 74.8 | 80.2 | 68.4 |
| A-GARCH(2,2) | 20.1 | 19.5 | 3.6 | 8.5 | 18.8 | 40.1 | 41.0 | 31.6 | 35.6 | 28.0 | 29.8 | 62.9 | 56.5 | 59.0 | 32.5 |
| NA-GARCH(1,1) | 56.5 | 68.7 | 45.6 | 75.7 | 72.9 | 71.7 | 77.2 | 49.5 | 59.9 | 38.9 | 49.8 | 73.6 | 64.1 | 67.2 | 62.2 |
| NA-GARCH(2,1) | 47.1 | 51.7 | 30.7 | 50.2 | 54.1 | 54.7 | 60.5 | 27.7 | 29.8 | 24.0 | 20.4 | 45.3 | 41.0 | 51.4 | 42.0 |
| NA-GARCH(1,2) | 87.5 | 82.1 | 23.4 | 69.3 | 84.5 | 93.0 | 92.1 | 48.6 | 57.8 | 35.0 | 45.6 | 74.8 | 67.5 | 70.5 | 66.5 |
| NA-GARCH(2,2) | 8.8 | 9.1 | .6 | 2.1 | 10.0 | 16.1 | 17.3 | 29.2 | 28.6 | 25.2 | 20.1 | 44.7 | 37.7 | 48.0 | 21.3 |
| V-GARCH(1,1) | 31.9 | 40.7 | 30.4 | 36.8 | 39.5 | 80.9 | 71.7 | 8.5 | 24.9 | 8.2 | 9.1 | 90.9 | 99.4 | 99.4 | 48.0 |
| V-GARCH(2,1) | 31.3 | 29.2 | 17.0 | 24.9 | 27.1 | 70.8 | 55.0 | 4.3 | 14.0 | 5.2 | 4.3 | 51.7 | 82.7 | 80.5 | 35.6 |
| V-GARCH(1,2) | 28.0 | 36.5 | 16.1 | 24.0 | 45.6 | 84.5 | 77.5 | 9.1 | 24.0 | 7.9 | 8.5 | 89.4 | 99.1 | 99.1 | 46.4 |
| V-GARCH(2,2) | 18.2 | 15.8 | 7.0 | 10.0 | 13.1 | 44.1 | 35.0 | 3.3 | 12.8 | 3.6 | 3.3 | 46.2 | 78.7 | 76.9 | 26.3 |
| THR-GARCH(1,1) | 25.2 | 27.7 | 75.4 | 37.4 | 35.6 | 19.8 | 23.1 | 69.9 | 60.5 | 66.3 | 64.4 | 28.6 | 22.8 | 28.3 | 41.8 |
| THR-GARCH(2,1) | 24.3 | 25.2 | 69.9 | 30.4 | 29.5 | 14.0 | 15.8 | 55.9 | 41.9 | 64.1 | 58.7 | 25.8 | 18.5 | 22.2 | 35.5 |
| THR-GARCH(1,2) | 91.2 | 86.6 | 73.3 | 84.2 | 80.9 | 66.9 | 62.3 | 71.4 | 61.1 | 70.5 | 72.6 | 28.0 | 22.2 | 28.0 | 64.2 |
| THR-GARCH(2,2) | 8.5 | 11.2 | 8.8 | 10.9 | 14.0 | 10.0 | 10.9 | 55.6 | 41.6 | 63.8 | 58.4 | 25.5 | 18.2 | 21.9 | 25.7 |
| GJR-GARCH(1,1) | 79.0 | 89.7 | 56.5 | 95.7 | 91.2 | 84.8 | 88.4 | 26.1 | 23.7 | 26.1 | 28.3 | 41.3 | 52.9 | 63.2 | 60.5 |
| GJR-GARCH(2,1) | 69.3 | 75.1 | 46.2 | 83.9 | 79.0 | 77.5 | 81.2 | 18.8 | 19.1 | 12.8 | 16.1 | 35.6 | 48.6 | 58.7 | 51.6 |
| GJR-GARCH(1,2) | 83.6 | 78.7 | 17.9 | 58.7 | 82.7 | 90.6 | 90.0 | 24.6 | 20.7 | 23.1 | 24.0 | 39.2 | 55.0 | 64.7 | 53.8 |
| GJR-GARCH(2,2) | 15.2 | 17.9 | 8.5 | 16.1 | 20.7 | 41.3 | 44.4 | 49.8 | 32.8 | 51.4 | 50.5 | 21.9 | 28.9 | 30.4 | 30.7 |
| LOG-GARCH(1,1) | 81.2 | 72.0 | 93.0 | 79.0 | 65.0 | 52.0 | 51.7 | 82.1 | 77.8 | 75.4 | 81.8 | 43.2 | 36.2 | 34.3 | 66.0 |
| LOG-GARCH(2,1) | 84.2 | 69.6 | 95.1 | 76.3 | 59.6 | 51.4 | 51.1 | 63.2 | 47.7 | 90.3 | 67.5 | 20.4 | 16.4 | 16.1 | 57.8 |
| LOG-GARCH(1,2) | 99.4 | 98.5 | 41.3 | 93.9 | 99.4 | 97.6 | 97.3 | 79.0 | 73.6 | 77.2 | 81.2 | 38.6 | 33.1 | 31.9 | 74.4 |
| LOG-GARCH(2,2) | **100.0** | **100.0** | 35.6 | 95.1 | 99.7 | 99.4 | 99.1 | 62.9 | 42.2 | 93.6 | 62.0 | 17.9 | 13.4 | 13.7 | 66.8 |
| EGARCH(1,1) | 37.1 | 38.6 | 71.1 | 40.7 | 38.6 | 38.6 | 36.5 | 70.8 | 76.6 | 53.5 | 63.2 | 60.2 | 56.8 | 55.3 | 52.7 |
| EGARCH(2,1) | 42.9 | 39.5 | 75.7 | 41.6 | 38.3 | 35.6 | 33.7 | 53.2 | 50.5 | 48.3 | 52.9 | 38.9 | 35.3 | 39.8 | 44.7 |
| EGARCH(1,2) | 99.7 | 99.7 | 45.3 | 95.4 | 98.5 | 97.9 | 97.6 | 68.7 | 73.3 | 55.6 | 69.6 | 55.0 | 52.6 | 53.2 | 75.9 |
| EGARCH(2,2) | 11.6 | 13.4 | 9.7 | 13.1 | 15.5 | 14.9 | 15.2 | 52.9 | 48.0 | 46.8 | 51.4 | 37.7 | 34.7 | 38.6 | 28.8 |
| NGARCH(1,1) | 83.0 | 91.2 | 86.6 | 97.9 | 92.4 | 72.3 | 79.0 | 96.7 | 67.5 | 97.6 | 90.9 | 21.0 | 17.6 | 15.5 | 72.1 |
| NGARCH(2,1) | 80.2 | 81.8 | 87.8 | 96.4 | 83.9 | 64.7 | 74.5 | 83.6 | 34.0 | **100.0** | 77.5 | 11.9 | 9.1 | 9.1 | 63.9 |
| NGARCH(1,2) | 92.7 | 94.8 | 38.9 | 88.8 | 95.1 | 88.4 | 87.2 | 97.3 | 67.2 | 97.9 | 92.7 | 19.5 | 14.9 | 14.6 | 70.7 |
| NGARCH(2,2) | 94.5 | 96.4 | 35.3 | 93.0 | 98.8 | 92.1 | 91.5 | 83.9 | 33.1 | 99.7 | 74.5 | 11.6 | 8.8 | 8.8 | 65.8 |
| A-PARCH(1,1) | 43.8 | 60.8 | 75.1 | 76.6 | 71.4 | 46.5 | 51.4 | 81.2 | 53.2 | 83.9 | 69.3 | 21.6 | 15.5 | 16.7 | 54.8 |
| A-PARCH(2,1) | 38.3 | 48.6 | 65.3 | 58.1 | 58.1 | 37.7 | 43.5 | 56.2 | 31.6 | 77.5 | 57.8 | 17.3 | 12.5 | 13.1 | 44.0 |
| A-PARCH(1,2) | 93.0 | 95.4 | 39.5 | 89.7 | 95.4 | 89.1 | 87.8 | 84.5 | 55.0 | 89.4 | 76.3 | 22.5 | 16.1 | 17.0 | 67.9 |
| A-PARCH(2,2) | 52.0 | 65.0 | 24.6 | 52.9 | 75.7 | 63.8 | 66.3 | 56.5 | 31.9 | 78.1 | 58.1 | 17.6 | 12.2 | 13.4 | 47.7 |
| GQ-ARCH(1,1) | 71.4 | 77.8 | 49.2 | 86.3 | 79.9 | 79.3 | 83.0 | 47.4 | 65.7 | 38.3 | 49.2 | 81.8 | 70.8 | 72.6 | 68.1 |
| GQ-ARCH(2,1) | 77.8 | 95.1 | 99.4 | 99.7 | 97.0 | 80.5 | 91.2 | 18.2 | 27.7 | 10.6 | 13.1 | 48.0 | 48.0 | 52.3 | 61.3 |
| GQ-ARCH(1,2) | 85.4 | 80.9 | 21.0 | 66.0 | 87.2 | 92.4 | 92.4 | 45.3 | 64.1 | 35.3 | 46.8 | 85.7 | 74.5 | 79.9 | 68.3 |
| GQ-ARCH(2,2) | 21.3 | 22.2 | 10.9 | 21.6 | 25.2 | 27.1 | 27.4 | 9.7 | 8.2 | 15.2 | 9.4 | 9.7 | 12.8 | 11.6 | 16.6 |
| H-GARCH(1,1) | 39.5 | 48.0 | 57.4 | 54.7 | 49.2 | 44.7 | 46.8 | 67.8 | 18.5 | 95.1 | 56.5 | 10.6 | 8.2 | 8.2 | 43.2 |
| AUG-GARCH(1,1) | 43.5 | 46.8 | 55.0 | 47.7 | 44.7 | 50.8 | 48.9 | 58.1 | 12.2 | 96.4 | 51.1 | 9.1 | 6.4 | 7.0 | 41.3 |

Relative performance ranking. Each row corresponds to a particular model, and a score shows the percentage of models (out of the total of 333) that performed worse than the particular model, measures 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),. . . , (11), are here denoted by $L_1, \ldots, L_7$. The last column is the average of the 14 scores.

Table 6: Models with Gaussian error distribution and constant mean

| Model | Exchange Rate Data | | | | | | | IBM Data | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | |
| ARCH(1) | 4.9 | 4.3 | .9 | 1.2 | 4.6 | 8.5 | 5.8 | 1.2 | 1.2 | 1.2 | 1.2 | 8.5 | 29.5 | 21.0 | 6.7 |
| GARCH(1,1) | 84.8 | 91.8 | 62.3 | 97.3 | 92.7 | 87.2 | 89.7 | 43.5 | 45.0 | 41.9 | 43.8 | 45.0 | 39.8 | 38.9 | 64.5 |
| GARCH(2,1) | 82.4 | 85.7 | 59.3 | 91.5 | 83.3 | 85.7 | 88.1 | 29.5 | 15.8 | 45.3 | 31.0 | 13.4 | 15.8 | 14.3 | 52.9 |
| GARCH(1,2) | 87.8 | 84.2 | 21.6 | 68.4 | 86.3 | 94.5 | 94.2 | 41.9 | 37.4 | 43.2 | 42.9 | 34.0 | 33.7 | 33.7 | 57.4 |
| GARCH(2,2) | 89.7 | 90.3 | 18.8 | 74.8 | 91.8 | 97.0 | 97.0 | 48.3 | 34.7 | 48.9 | 43.2 | 22.2 | 27.4 | 26.1 | 57.9 |
| IGARCH(1,1) | 7.0 | 7.3 | 52.3 | 16.7 | 8.2 | 5.5 | 7.6 | 14.0 | 4.9 | 85.4 | 17.3 | 2.7 | 3.3 | 3.0 | 16.8 |
| IGARCH(2,1) | 5.8 | 6.4 | 44.7 | 15.2 | 7.0 | 4.9 | 7.0 | 16.7 | 7.3 | 73.3 | 19.8 | 5.8 | 5.8 | 5.8 | 16.1 |
| IGARCH(1,2) | 2.7 | 5.5 | 31.3 | 12.2 | 6.4 | 3.6 | 4.6 | 14.9 | 5.8 | 82.4 | 16.7 | 3.6 | 4.3 | 3.6 | 14.1 |
| IGARCH(2,2) | 10.3 | 8.5 | 41.0 | 11.2 | 5.8 | 7.6 | 8.5 | 40.4 | 9.1 | 84.2 | 37.7 | 7.6 | 7.3 | 6.7 | 20.4 |
| TS-GARCH(1,1) | 45.0 | 48.9 | 90.9 | 59.9 | 49.5 | 29.5 | 32.5 | 86.6 | 70.2 | 90.9 | 86.0 | 32.5 | 24.6 | 25.5 | 55.2 |
| TS-GARCH(2,1) | 46.5 | 47.4 | 96.0 | 60.8 | 46.8 | 28.6 | 32.2 | 72.9 | 69.0 | 79.3 | 85.1 | 38.3 | 32.5 | 33.4 | 54.9 |
| TS-GARCH(1,2) | 90.9 | 86.3 | 60.5 | 81.8 | 80.5 | 76.3 | 71.4 | 86.9 | 72.6 | 90.0 | 91.2 | 35.0 | 28.0 | 29.8 | 70.1 |
| TS-GARCH(2,2) | 94.2 | 92.4 | 54.7 | 87.8 | 89.4 | 83.3 | 81.5 | 84.2 | 75.1 | 84.8 | 91.8 | 40.1 | 32.8 | 33.1 | 73.2 |
| A-GARCH(1,1) | 71.1 | 77.5 | 47.1 | 85.4 | 78.7 | 80.2 | 83.6 | 45.9 | 64.7 | 37.1 | 48.6 | 83.0 | 71.1 | 72.9 | 67.7 |
| A-GARCH(2,1) | 60.2 | 59.6 | 28.3 | 53.8 | 62.0 | 68.4 | 75.4 | 36.8 | 39.5 | 28.6 | 30.4 | 67.2 | 59.9 | 64.4 | 52.5 |
| A-GARCH(1,2) | 86.3 | 83.0 | 20.1 | 67.2 | 88.4 | 93.3 | 95.1 | 43.8 | 63.2 | 32.2 | 45.0 | 87.2 | 76.3 | 82.1 | 68.8 |
| A-GARCH(2,2) | 20.4 | 19.8 | 3.0 | 8.2 | 18.5 | 40.7 | 41.6 | 31.9 | 36.2 | 27.4 | 29.2 | 63.8 | 58.7 | 62.0 | 33.0 |
| NA-GARCH(1,1) | 58.1 | 69.3 | 42.2 | 73.9 | 73.3 | 74.8 | 79.3 | 49.2 | 60.2 | 38.0 | 48.9 | 74.2 | 65.0 | 68.4 | 62.5 |
| NA-GARCH(2,1) | 48.9 | 53.5 | 30.1 | 50.5 | 57.1 | 55.6 | 63.2 | 27.1 | 29.5 | 21.3 | 19.5 | 46.5 | 43.2 | 52.0 | 42.7 |
| NA-GARCH(1,2) | 88.4 | 85.1 | 22.5 | 69.0 | 85.4 | 94.8 | 93.6 | 48.0 | 58.1 | 32.8 | 44.4 | 76.0 | 68.4 | 71.4 | 67.0 |
| NA-GARCH(2,2) | 9.7 | 9.7 | .3 | 2.4 | 10.6 | 17.9 | 18.5 | 28.3 | 28.9 | 24.6 | 19.1 | 45.9 | 38.9 | 48.9 | 21.7 |
| V-GARCH(1,1) | 30.4 | 40.1 | 28.9 | 35.9 | 40.1 | 81.8 | 73.9 | 7.3 | 23.1 | 7.0 | 8.2 | 89.1 | 98.8 | 98.8 | 47.4 |
| V-GARCH(2,1) | 30.7 | 29.5 | 16.7 | 24.6 | 28.0 | 73.6 | 55.9 | 4.0 | 13.7 | 4.0 | 4.0 | 50.5 | 83.0 | 80.9 | 35.6 |
| V-GARCH(1,2) | 27.1 | 37.1 | 15.8 | 23.7 | 46.5 | 86.3 | 78.4 | 7.9 | 21.9 | 6.7 | 7.6 | 88.1 | 98.5 | 98.5 | 46.0 |
| V-GARCH(2,2) | 17.9 | 15.5 | 6.4 | 9.7 | 13.7 | 44.4 | 35.9 | 3.0 | 12.5 | 3.0 | 3.0 | 45.6 | 79.0 | 77.5 | 26.2 |
| THR-GARCH(1,1) | 24.9 | 27.1 | 69.6 | 35.0 | 33.1 | 19.5 | 21.9 | 70.2 | 61.4 | 67.2 | 66.3 | 30.1 | 23.4 | 29.5 | 41.4 |
| THR-GARCH(2,1) | 24.6 | 24.6 | 66.9 | 29.5 | 27.7 | 14.3 | 16.1 | 57.1 | 43.5 | 63.5 | 59.6 | 26.7 | 19.1 | 24.0 | 35.5 |
| THR-GARCH(1,2) | 90.3 | 84.8 | 64.1 | 80.9 | 77.5 | 69.9 | 66.9 | 71.7 | 64.4 | 69.9 | 75.4 | 30.4 | 23.1 | 29.2 | 64.2 |
| THR-GARCH(2,2) | 10.0 | 11.6 | 8.2 | 10.6 | 12.2 | 10.6 | 11.9 | 56.8 | 43.2 | 63.2 | 59.0 | 26.4 | 18.8 | 23.7 | 26.2 |
| GJR-GARCH(1,1) | 77.5 | 87.8 | 53.8 | 94.5 | 90.0 | 83.9 | 86.9 | 26.7 | 25.8 | 25.5 | 28.0 | 41.9 | 55.3 | 63.5 | 60.1 |
| GJR-GARCH(2,1) | 69.6 | 74.2 | 44.1 | 83.0 | 76.9 | 78.1 | 80.9 | 21.6 | 20.1 | 12.2 | 16.4 | 36.5 | 51.4 | 60.5 | 51.8 |
| GJR-GARCH(1,2) | 82.1 | 76.0 | 17.3 | 56.8 | 82.1 | 90.9 | 90.3 | 25.2 | 21.3 | 19.1 | 23.1 | 39.8 | 56.2 | 65.3 | 53.3 |
| GJR-GARCH(2,2) | 16.1 | 17.6 | 7.6 | 15.8 | 20.4 | 42.6 | 45.0 | 50.2 | 34.3 | 49.5 | 50.2 | 24.0 | 30.7 | 32.2 | 31.2 |
| LOG-GARCH(1,1) | 72.3 | 64.1 | 89.1 | 65.0 | 56.2 | 48.6 | 45.6 | 82.7 | 78.7 | 73.6 | 83.6 | 47.4 | 38.3 | 35.9 | 62.9 |
| LOG-GARCH(2,1) | 78.4 | 62.9 | 91.8 | 63.2 | 51.1 | 48.3 | 46.2 | 66.9 | 61.7 | 88.1 | 76.9 | 26.1 | 21.3 | 21.3 | 57.4 |
| LOG-GARCH(1,2) | 98.2 | 98.2 | 36.5 | 90.6 | 96.7 | 99.1 | 98.8 | 80.2 | 76.9 | 75.7 | 83.9 | 42.9 | 34.0 | 34.0 | 74.7 |
| LOG-GARCH(2,2) | 99.1 | 99.4 | 33.7 | 92.1 | 99.1 | 99.7 | 99.4 | 67.5 | 51.7 | 90.6 | 68.7 | 20.1 | 20.1 | 18.2 | 68.5 |
| EGARCH(1,1) | 36.8 | 38.0 | 67.2 | 38.9 | 37.1 | 37.4 | 35.6 | 70.5 | 77.2 | 53.8 | 64.1 | 61.4 | 57.8 | 55.6 | 52.2 |
| EGARCH(2,1) | 44.1 | 39.2 | 74.8 | 40.1 | 36.8 | 37.1 | 34.3 | 54.4 | 53.8 | 47.7 | 54.7 | 42.6 | 37.1 | 41.9 | 45.6 |
| EGARCH(1,2) | 97.3 | 97.6 | 34.7 | 89.4 | 96.0 | 98.5 | 98.2 | 69.0 | 74.5 | 55.0 | 70.2 | 58.4 | 53.8 | 53.8 | 74.7 |
| EGARCH(2,2) | 11.9 | 13.7 | 9.4 | 12.8 | 15.2 | 18.2 | 17.6 | 53.8 | 52.9 | 46.2 | 53.5 | 40.7 | 36.5 | 40.7 | 30.2 |
| NGARCH(1,1) | 76.6 | 85.4 | 83.6 | 96.7 | 85.7 | 61.1 | 69.6 | 98.8 | 70.8 | 96.0 | 89.4 | 24.3 | 24.3 | 23.4 | 70.4 |
| NGARCH(2,1) | 73.9 | 78.4 | 85.4 | 92.4 | 81.2 | 56.8 | 65.7 | 94.8 | 45.9 | 99.4 | 86.6 | 14.9 | 11.9 | 10.9 | 64.2 |
| NGARCH(1,2) | 93.6 | 94.5 | 40.1 | 88.4 | 94.5 | 88.1 | 86.0 | 98.5 | 69.6 | 97.3 | 92.4 | 23.1 | 20.7 | 19.8 | 71.9 |
| NGARCH(2,2) | 95.4 | 96.7 | 35.9 | 93.3 | 97.9 | 96.0 | 93.3 | 94.5 | 45.3 | 98.8 | 81.5 | 14.0 | 11.6 | 11.2 | 69.0 |
| A-PARCH(1,1) | 41.6 | 58.4 | 71.4 | 71.1 | 66.3 | 45.3 | 49.5 | 83.3 | 52.3 | 86.6 | 68.1 | 19.8 | 14.6 | 15.8 | 53.1 |
| A-PARCH(2,1) | 38.6 | 47.7 | 64.7 | 57.1 | 55.6 | 38.3 | 43.2 | 55.3 | 22.5 | 74.5 | 54.4 | 15.5 | 10.9 | 11.9 | 42.2 |
| A-PARCH(1,2) | 92.4 | 93.3 | 37.4 | 87.2 | 93.6 | 88.8 | 86.3 | 85.1 | 52.0 | 93.0 | 76.6 | 19.1 | 13.1 | 15.2 | 66.7 |
| A-PARCH(2,2) | 49.8 | 63.2 | 23.1 | 51.1 | 74.8 | 64.1 | 67.2 | 55.0 | 21.6 | 71.4 | 52.0 | 15.8 | 11.2 | 12.5 | 45.2 |
| GQ-ARCH(1,1) | 70.8 | 77.2 | 46.8 | 85.1 | 78.4 | 79.9 | 83.3 | 45.6 | 65.3 | 36.8 | 48.3 | 83.3 | 71.4 | 73.3 | 67.5 |
| GQ-ARCH(2,1) | 69.0 | 91.5 | 98.8 | 99.4 | 94.8 | 77.8 | 89.4 | 18.5 | 28.3 | 10.0 | 12.8 | 49.2 | 50.2 | 53.5 | 60.2 |
| GQ-ARCH(1,2) | 86.6 | 83.3 | 20.4 | 67.5 | 88.8 | 93.6 | 95.4 | 44.1 | 63.5 | 32.5 | 45.9 | 87.5 | 76.6 | 82.4 | 69.1 |
| GQ-ARCH(2,2) | 21.6 | 22.8 | 11.2 | 22.2 | 26.1 | 27.7 | 29.2 | 10.3 | 8.5 | 13.1 | 9.7 | 10.3 | 13.7 | 12.2 | 17.0 |
| H-GARCH(1,1) | 42.6 | 52.3 | 56.2 | 55.6 | 52.0 | 48.0 | 50.5 | 71.1 | 22.8 | 91.2 | 57.1 | 12.2 | 9.4 | 9.4 | 45.0 |
| AUG-GARCH(1,1) | 41.3 | 45.0 | 51.7 | 45.6 | 42.9 | 50.2 | 47.7 | 59.6 | 15.5 | 95.7 | 52.6 | 10.0 | 7.6 | 7.9 | 40.9 |

Relative performance ranking. Each row corresponds to a particular model, and a score shows the percentage of models (out of the total of 333) that performed worse than the particular model, measures 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),. . . , (11), are here denoted by $L_1, \ldots, L_7$. The last column is the average of the 14 scores.

Table 7: Models with Gaussian error distribution and GARCH-in-mean

| Model | Exchange Rate Data | | | | | | | IBM Data | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | |
| ARCH(1) | 5.2 | 4.6 | 1.5 | 1.5 | 4.9 | 8.8 | 6.7 | .9 | .9 | .9 | .9 | 7.9 | 29.2 | 20.4 | 6.7 |
| GARCH(1,1) | 81.8 | 90.6 | 62.9 | 97.0 | 90.6 | 85.1 | 88.8 | 46.2 | 45.6 | 41.0 | 46.2 | 49.8 | 43.5 | 44.1 | 65.2 |
| GARCH(2,1) | 81.5 | 83.9 | 59.6 | 90.9 | 83.0 | 84.2 | 86.6 | 30.4 | 16.4 | 45.6 | 31.6 | 14.3 | 17.0 | 14.9 | 52.8 |
| GARCH(1,2) | 88.1 | 83.6 | 20.7 | 68.1 | 86.0 | 95.4 | 94.5 | 42.9 | 41.3 | 42.6 | 43.5 | 37.1 | 34.3 | 34.7 | 58.1 |
| GARCH(2,2) | 90.0 | 90.0 | 18.5 | 74.2 | 91.5 | 96.7 | 96.7 | 50.5 | 35.3 | 51.1 | 47.1 | 23.4 | 30.1 | 28.6 | 58.8 |
| IGARCH(1,1) | 6.4 | 7.0 | 51.4 | 16.4 | 7.6 | 4.3 | 5.2 | 14.6 | 5.5 | 84.5 | 18.5 | 3.3 | 4.0 | 4.0 | 16.6 |
| IGARCH(2,1) | 6.1 | 6.7 | 44.4 | 15.5 | 7.3 | 5.2 | 7.3 | 17.3 | 7.9 | 72.0 | 21.6 | 6.1 | 6.1 | 6.1 | 16.4 |
| IGARCH(1,2) | 3.0 | 5.8 | 31.0 | 12.5 | 6.7 | 4.0 | 4.9 | 15.2 | 6.1 | 81.2 | 17.6 | 4.0 | 4.9 | 4.6 | 14.4 |
| IGARCH(2,2) | 8.2 | 8.2 | 38.0 | 9.1 | 5.5 | 6.4 | 6.1 | 42.2 | 10.3 | 86.9 | 41.9 | 8.8 | 7.9 | 7.3 | 20.5 |
| TS-GARCH(1,1) | 45.6 | 49.2 | 91.5 | 60.2 | 49.8 | 30.1 | 33.4 | 83.0 | 67.8 | 91.5 | 82.4 | 29.8 | 21.9 | 23.1 | 54.2 |
| TS-GARCH(2,1) | 48.3 | 48.3 | 96.7 | 62.0 | 47.4 | 29.2 | 32.8 | 72.3 | 68.4 | 78.7 | 84.8 | 37.4 | 31.9 | 32.5 | 55.1 |
| TS-GARCH(1,2) | 90.6 | 86.0 | 62.0 | 82.1 | 80.2 | 76.0 | 70.8 | 84.8 | 69.3 | 89.7 | 87.5 | 33.4 | 25.5 | 26.7 | 68.9 |
| TS-GARCH(2,2) | 93.9 | 92.1 | 54.4 | 87.5 | 89.1 | 83.0 | 80.5 | 80.5 | 70.5 | 83.3 | 88.1 | 38.0 | 31.0 | 31.3 | 71.7 |
| A-GARCH(1,1) | 68.7 | 76.6 | 48.0 | 84.8 | 78.1 | 79.0 | 82.4 | 46.5 | 62.9 | 37.4 | 47.4 | 80.9 | 70.2 | 72.0 | 66.8 |
| A-GARCH(2,1) | 61.4 | 60.2 | 27.4 | 53.5 | 62.6 | 69.3 | 76.6 | 37.4 | 41.0 | 29.2 | 31.3 | 67.5 | 59.6 | 63.8 | 52.9 |
| A-GARCH(1,2) | 86.9 | 82.4 | 19.5 | 66.6 | 87.8 | 93.9 | 94.8 | 44.7 | 62.3 | 33.7 | 45.3 | 85.1 | 73.6 | 79.0 | 68.3 |
| A-GARCH(2,2) | 19.8 | 20.1 | 4.0 | 8.8 | 19.1 | 40.4 | 41.9 | 35.6 | 38.3 | 27.7 | 30.1 | 64.7 | 58.1 | 61.1 | 33.5 |
| NA-GARCH(1,1) | 56.8 | 69.0 | 42.9 | 74.5 | 73.6 | 73.9 | 78.7 | 48.9 | 59.3 | 36.5 | 47.7 | 73.9 | 64.4 | 68.1 | 62.0 |
| NA-GARCH(2,1) | 50.8 | 54.1 | 28.6 | 49.8 | 54.4 | 56.2 | 63.5 | 25.8 | 27.4 | 21.6 | 18.8 | 44.1 | 41.3 | 51.7 | 42.0 |
| NA-GARCH(1,2) | 88.8 | 84.5 | 22.2 | 68.7 | 85.1 | 95.7 | 93.9 | 47.7 | 56.8 | 31.9 | 44.1 | 75.4 | 67.8 | 71.1 | 66.7 |
| NA-GARCH(2,2) | 9.1 | 9.4 | .0 | 1.8 | 10.3 | 16.7 | 17.9 | 27.4 | 27.1 | 24.3 | 17.9 | 44.4 | 38.6 | 48.3 | 21.0 |
| V-GARCH(1,1) | 30.1 | 40.4 | 29.5 | 36.5 | 40.4 | 81.2 | 73.6 | 8.2 | 24.6 | 7.3 | 8.8 | 91.2 | **100.0** | 99.7 | 48.0 |
| V-GARCH(2,1) | 31.6 | 29.8 | 16.4 | 24.3 | 28.3 | 74.2 | 56.2 | 4.6 | 14.9 | 4.9 | 4.6 | 53.2 | 84.8 | 82.7 | 36.5 |
| V-GARCH(1,2) | 26.4 | 36.8 | 15.5 | 23.4 | 46.2 | 86.0 | 78.1 | 8.8 | 24.3 | 7.6 | 7.9 | 90.6 | 99.7 | **100.0** | 46.5 |
| V-GARCH(2,2) | 17.6 | 15.2 | 6.1 | 9.4 | 12.8 | 43.8 | 35.3 | 3.6 | 13.1 | 3.3 | 3.6 | 47.7 | 81.5 | 79.3 | 26.6 |
| THR-GARCH(1,1) | 28.3 | 28.0 | 69.3 | 36.2 | 34.3 | 21.6 | 23.7 | 68.4 | 59.6 | 66.6 | 64.7 | 29.2 | 22.5 | 27.7 | 41.4 |
| THR-GARCH(2,1) | 27.7 | 25.8 | 68.4 | 31.3 | 29.2 | 15.8 | 19.8 | 57.8 | 44.4 | 62.9 | 59.9 | 27.4 | 20.4 | 25.2 | 36.8 |
| THR-GARCH(1,2) | 89.4 | 80.2 | 66.0 | 79.3 | 76.3 | 65.7 | 61.7 | 69.6 | 60.8 | 69.6 | 72.3 | 28.9 | 21.6 | 27.4 | 62.0 |
| THR-GARCH(2,2) | 9.4 | 10.9 | 7.9 | 10.3 | 11.9 | 10.3 | 11.2 | 57.4 | 43.8 | 62.6 | 59.3 | 27.1 | 19.8 | 24.6 | 26.2 |
| GJR-GARCH(1,1) | 75.1 | 86.9 | 55.3 | 94.2 | 89.7 | 82.1 | 85.1 | 26.4 | 25.5 | 24.9 | 26.7 | 42.2 | 55.9 | 64.1 | 59.6 |
| GJR-GARCH(2,1) | 70.5 | 74.5 | 43.8 | 82.7 | 77.2 | 78.4 | 82.1 | 20.4 | 19.8 | 11.2 | 15.8 | 36.2 | 52.3 | 61.7 | 51.9 |
| GJR-GARCH(1,2) | 82.7 | 76.9 | 17.6 | 57.4 | 82.4 | 91.2 | 90.6 | 24.9 | 21.0 | 18.8 | 22.5 | 39.5 | 57.1 | 65.7 | 53.5 |
| GJR-GARCH(2,2) | 15.5 | 17.3 | 7.3 | 14.9 | 19.5 | 42.9 | 45.3 | 50.8 | 33.7 | 50.5 | 50.8 | 22.8 | 30.4 | 30.7 | 30.9 |
| LOG-GARCH(1,1) | 73.3 | 64.4 | 88.8 | 64.1 | 55.9 | 48.9 | 45.9 | 79.9 | 77.5 | 72.3 | 80.5 | 43.8 | 35.6 | 35.0 | 61.9 |
| LOG-GARCH(2,1) | 79.3 | 63.8 | 92.1 | 64.7 | 51.7 | 49.5 | 47.1 | 65.3 | 55.9 | 88.4 | 75.1 | 24.9 | 21.0 | 20.1 | 57.1 |
| LOG-GARCH(1,2) | 97.6 | 97.9 | 36.8 | 90.3 | 96.4 | 98.8 | 98.5 | 77.5 | 72.0 | 75.1 | 80.9 | 41.0 | 33.4 | 32.8 | 73.5 |
| LOG-GARCH(2,2) | 98.8 | 99.1 | 34.0 | 91.8 | 98.2 | **100.0** | 99.7 | 67.2 | 46.5 | 93.3 | 67.8 | 18.5 | 17.3 | 16.4 | 67.8 |
| EGARCH(1,1) | 38.0 | 38.3 | 67.5 | 39.2 | 38.0 | 39.5 | 37.7 | 69.3 | 74.8 | 53.2 | 62.9 | 59.6 | 55.6 | 55.0 | 52.0 |
| EGARCH(2,1) | 48.6 | 41.6 | 74.2 | 42.9 | 37.7 | 38.9 | 37.4 | 54.1 | 53.5 | 48.0 | 54.1 | 41.6 | 36.8 | 41.0 | 46.5 |
| EGARCH(1,2) | 97.0 | 97.3 | 34.3 | 89.1 | 95.7 | 98.2 | 97.9 | 68.1 | 72.3 | 54.7 | 69.0 | 55.3 | 51.7 | 52.6 | 73.8 |
| EGARCH(2,2) | 11.2 | 13.1 | 9.1 | 11.9 | 14.6 | 17.6 | 17.0 | 53.5 | 52.6 | 46.5 | 53.2 | 40.4 | 35.9 | 40.1 | 29.8 |
| NGARCH(1,1) | 74.5 | 81.5 | 83.3 | 96.0 | 84.8 | 60.2 | 68.7 | 91.5 | 66.0 | 96.7 | 84.2 | 20.7 | 19.5 | 19.1 | 67.6 |
| NGARCH(2,1) | 74.2 | 79.0 | 85.1 | 92.7 | 81.5 | 57.1 | 66.6 | 88.4 | 44.7 | 98.5 | 82.1 | 14.6 | 10.6 | 10.6 | 63.3 |
| NGARCH(1,2) | 93.3 | 94.2 | 39.8 | 88.1 | 94.2 | 87.8 | 85.7 | 85.7 | 39.8 | 98.2 | 77.8 | 13.1 | 10.3 | 10.0 | 65.6 |
| NGARCH(2,2) | 95.1 | 97.0 | 36.2 | 93.6 | 97.6 | 95.1 | 93.0 | 85.4 | 35.0 | 99.1 | 77.2 | 12.8 | 10.0 | 9.7 | 66.9 |
| A-PARCH(1,1) | 43.2 | 59.0 | 70.5 | 71.7 | 66.6 | 46.2 | 50.2 | 72.0 | 49.5 | 67.8 | 60.5 | 24.6 | 16.7 | 19.5 | 51.3 |
| A-PARCH(2,1) | 40.7 | 49.8 | 63.8 | 57.8 | 56.8 | 39.8 | 44.1 | 58.4 | 35.9 | 64.4 | 56.8 | 21.3 | 14.0 | 17.9 | 44.4 |
| A-PARCH(1,2) | 92.1 | 93.0 | 37.1 | 86.9 | 93.9 | 89.4 | 87.5 | 76.6 | 50.8 | 76.0 | 63.5 | 23.7 | 15.2 | 18.8 | 64.6 |
| A-PARCH(2,2) | 49.2 | 61.7 | 23.7 | 50.8 | 74.2 | 61.7 | 65.3 | 54.7 | 19.5 | 76.6 | 55.3 | 13.7 | 9.7 | 10.3 | 44.7 |
| GQ-ARCH(1,1) | 68.4 | 76.3 | 48.3 | 84.5 | 77.8 | 78.7 | 81.8 | 46.8 | 62.6 | 37.7 | 48.0 | 80.5 | 69.9 | 71.7 | 66.7 |
| GQ-ARCH(2,1) | 83.3 | 96.0 | 98.5 | **100.0** | 97.3 | 83.6 | 96.0 | 21.3 | 30.4 | 10.3 | 14.0 | 50.2 | 51.1 | 54.1 | 63.3 |
| GQ-ARCH(1,2) | 87.2 | 82.7 | 19.1 | 66.9 | 88.1 | 94.2 | 95.7 | 44.4 | 62.0 | 33.4 | 44.7 | 84.8 | 73.9 | 79.6 | 68.3 |
| GQ-ARCH(2,2) | 20.7 | 21.9 | 11.6 | 21.9 | 25.5 | 25.8 | 27.1 | 14.3 | 10.0 | 25.8 | 11.2 | 10.9 | 17.9 | 14.0 | 18.5 |
| H-GARCH(1,1) | 40.4 | 42.6 | 52.0 | 41.9 | 40.7 | 43.5 | 43.8 | 66.3 | 18.8 | 95.4 | 57.4 | 11.2 | 8.5 | 8.5 | 40.8 |
| AUG-GARCH(1,1) | 44.7 | 45.3 | 48.6 | 43.8 | 42.2 | 50.5 | 48.0 | 58.7 | 14.6 | 97.0 | 55.0 | 9.4 | 6.7 | 7.6 | 40.9 |

Relative performance ranking. Each row corresponds to a particular model, and a score shows the percentage of models (out of the total of 333) that performed worse than the particular model, measures 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),. . . , (11), are here denoted by $L_1, \ldots, L_7$. The last column is the average of the 14 scores.

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

| Model | Exchange Rate Data | | | | | | | IBM Data | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | |
| ARCH(1) | 3.3 | 1.8 | 2.1 | .0 | .6 | 6.7 | 3.6 | .3 | .3 | .6 | .6 | 6.4 | 24.9 | 17.6 | 4.9 |
| GARCH(1,1) | 77.2 | 73.9 | 77.8 | 83.3 | 72.0 | 69.0 | 71.1 | 19.1 | 23.4 | 31.6 | 20.7 | 31.0 | 40.1 | 36.8 | 51.9 |
| GARCH(2,1) | 79.6 | 73.3 | 77.5 | 82.4 | 69.6 | 74.5 | 75.7 | 19.5 | 20.4 | 34.3 | 23.7 | 27.7 | 37.4 | 35.6 | 52.2 |
| GARCH(1,2) | 72.0 | 62.0 | 38.6 | 53.2 | 60.5 | 71.4 | 69.9 | 19.8 | 25.2 | 31.0 | 21.3 | 31.3 | 39.2 | 36.5 | 45.1 |
| GARCH(2,2) | 80.9 | 75.4 | 32.5 | 63.8 | 75.4 | 85.4 | 84.8 | 28.6 | 30.1 | 36.2 | 27.7 | 30.7 | 39.5 | 37.1 | 52.0 |
| IGARCH(1,1) | 2.4 | 3.6 | 59.0 | 6.4 | 3.3 | 1.8 | 2.1 | 9.4 | 1.8 | 94.2 | 11.9 | .0 | .0 | .0 | 14.0 |
| IGARCH(2,1) | 2.1 | 3.3 | 58.4 | 6.7 | 3.6 | 2.4 | 2.7 | 12.2 | 4.0 | 82.1 | 15.2 | 2.4 | 1.8 | 1.8 | 14.2 |
| IGARCH(1,2) | 1.8 | 3.0 | 55.6 | 5.8 | 3.0 | 2.1 | 2.4 | 10.6 | 2.4 | 91.8 | 12.5 | .9 | .6 | .6 | 13.8 |
| IGARCH(2,2) | 7.9 | 5.2 | 58.1 | 7.9 | 5.2 | 3.3 | 3.3 | 16.4 | 6.4 | 85.7 | 21.0 | 4.6 | 3.6 | 4.3 | 16.6 |
| TS-GARCH(1,1) | 34.7 | 32.5 | 97.0 | 41.0 | 34.7 | 15.2 | 15.5 | 78.4 | 80.2 | 69.0 | 83.0 | 53.8 | 47.7 | 45.3 | 52.0 |
| TS-GARCH(2,1) | 35.9 | 34.0 | 98.2 | 43.5 | 35.3 | 17.3 | 18.2 | 74.2 | 78.1 | 74.2 | 86.9 | 48.3 | 42.6 | 39.2 | 51.8 |
| TS-GARCH(1,2) | 47.7 | 44.7 | 82.4 | 47.4 | 44.1 | 30.4 | 30.1 | 79.6 | 81.5 | 70.2 | 87.8 | 56.5 | 45.9 | 42.6 | 56.5 |
| TS-GARCH(2,2) | 59.6 | 67.2 | 70.8 | 71.4 | 71.7 | 47.7 | 50.8 | 75.1 | 78.4 | 73.9 | 87.2 | 48.6 | 42.9 | 39.5 | 63.2 |
| A-GARCH(1,1) | 72.9 | 70.2 | 74.5 | 78.1 | 67.5 | 66.0 | 67.5 | 34.7 | 49.8 | 21.9 | 33.4 | 84.2 | 86.0 | 86.6 | 63.8 |
| A-GARCH(2,1) | 60.8 | 54.7 | 38.3 | 49.2 | 48.6 | 58.1 | 59.9 | 34.3 | 42.9 | 29.8 | 38.6 | 55.6 | 62.6 | 57.8 | 49.4 |
| A-GARCH(1,2) | 58.4 | 50.2 | 26.1 | 41.3 | 55.3 | 62.3 | 64.4 | 33.1 | 47.1 | 23.7 | 36.5 | 77.2 | 77.5 | 77.8 | 52.2 |
| A-GARCH(2,2) | 19.5 | 18.5 | 5.2 | 11.6 | 16.1 | 31.6 | 31.3 | 52.3 | 71.7 | 39.8 | 56.2 | 77.8 | 83.6 | 76.6 | 42.3 |
| NA-GARCH(1,1) | 66.9 | 66.9 | 71.7 | 72.0 | 64.7 | 60.5 | 62.6 | 41.3 | 58.7 | 17.9 | 39.5 | 96.0 | 91.2 | 93.0 | 64.5 |
| NA-GARCH(2,1) | 56.2 | 51.4 | 40.4 | 49.5 | 47.7 | 53.8 | 56.5 | 41.6 | 59.0 | 18.2 | 40.4 | 95.7 | 90.6 | 92.4 | 56.7 |
| NA-GARCH(1,2) | 61.1 | 55.6 | 25.5 | 44.1 | 60.8 | 69.6 | 70.2 | 39.2 | 55.6 | 17.0 | 41.3 | 93.9 | 88.1 | 89.1 | 57.9 |
| NA-GARCH(2,2) | 12.8 | 12.5 | 3.3 | 7.0 | 11.6 | 21.9 | 24.3 | 38.3 | 54.1 | 17.6 | 41.0 | 87.8 | 86.9 | 87.8 | 36.2 |
| V-GARCH(1,1) | 29.8 | 34.7 | 41.9 | 30.1 | 30.4 | 67.2 | 52.9 | 6.1 | 18.2 | 6.1 | 6.7 | 72.6 | 97.9 | 97.9 | 42.3 |
| V-GARCH(2,1) | 34.0 | 34.3 | 27.1 | 27.1 | 30.1 | 77.2 | 57.8 | 1.8 | 11.2 | 2.1 | 1.8 | 48.9 | 93.3 | 91.5 | 38.4 |
| V-GARCH(1,2) | 22.2 | 21.0 | 14.9 | 20.4 | 19.8 | 41.0 | 28.6 | 5.2 | 17.3 | 4.6 | 5.5 | 71.1 | 97.3 | 97.3 | 33.3 |
| V-GARCH(2,2) | 13.7 | 10.0 | 4.9 | 4.3 | 9.1 | 21.0 | 12.5 | 11.9 | 26.4 | 8.5 | 10.0 | 68.4 | 94.8 | 94.8 | 27.9 |
| THR-GARCH(1,1) | 23.7 | 24.0 | 87.5 | 27.7 | 24.0 | 11.9 | 11.6 | 77.2 | 88.8 | 51.7 | 74.2 | 88.4 | 75.4 | 74.2 | 52.9 |
| THR-GARCH(2,1) | 23.1 | 23.1 | 79.9 | 25.2 | 21.9 | 9.7 | 9.4 | 82.4 | 83.6 | 68.4 | 94.2 | 63.2 | 48.3 | 50.2 | 48.8 |
| THR-GARCH(1,2) | 48.0 | 46.2 | 76.9 | 47.1 | 45.3 | 38.0 | 39.2 | 74.8 | 89.4 | 52.9 | 79.0 | 90.0 | 75.7 | 75.7 | 62.7 |
| THR-GARCH(2,2) | 12.5 | 14.3 | 13.4 | 19.5 | 17.0 | 11.2 | 13.4 | 99.4 | 99.7 | 76.9 | 99.4 | 92.7 | 83.3 | 76.3 | 52.1 |
| GJR-GARCH(1,1) | 67.8 | 69.9 | 69.0 | 78.4 | 70.2 | 65.0 | 67.8 | 24.0 | 36.5 | 16.1 | 26.1 | 68.1 | 80.2 | 84.2 | 58.8 |
| GJR-GARCH(2,1) | 63.2 | 61.1 | 57.1 | 62.3 | 59.9 | 59.6 | 64.1 | 28.0 | 39.2 | 27.1 | 34.7 | 51.1 | 62.0 | 58.4 | 52.0 |
| GJR-GARCH(1,2) | 47.4 | 41.0 | 24.0 | 33.7 | 43.8 | 56.5 | 58.1 | 22.5 | 37.7 | 13.7 | 27.4 | 69.0 | 79.9 | 83.9 | 45.6 |
| GJR-GARCH(2,2) | 16.7 | 17.0 | 10.0 | 18.8 | 16.7 | 31.9 | 36.2 | 51.1 | 66.3 | 40.4 | 51.7 | 64.4 | 65.3 | 66.3 | 39.5 |
| LOG-GARCH(1,1) | 55.9 | 42.2 | 94.8 | 44.7 | 36.2 | 29.8 | 25.5 | 66.6 | 82.4 | 59.6 | 73.9 | 71.7 | 67.2 | 61.4 | 58.0 |
| LOG-GARCH(2,1) | 66.0 | 45.6 | 97.6 | 48.0 | 37.4 | 33.4 | 30.7 | 59.9 | 74.2 | 59.0 | 65.3 | 61.1 | 60.5 | 54.7 | 56.7 |
| LOG-GARCH(1,2) | 96.4 | 87.2 | 65.0 | 67.8 | 71.1 | 82.7 | 72.3 | 64.7 | 81.2 | 58.7 | 72.9 | 72.3 | 64.7 | 57.4 | 72.5 |
| LOG-GARCH(2,2) | 97.9 | 92.7 | 59.9 | 81.5 | 79.3 | 90.0 | 79.6 | 60.5 | 73.9 | 59.3 | 65.0 | 60.5 | 60.2 | 54.4 | 72.5 |
| EGARCH(1,1) | 35.6 | 30.7 | 84.8 | 33.4 | 28.6 | 24.0 | 21.6 | 65.0 | 90.6 | 44.1 | 61.7 | 99.1 | 95.7 | 95.7 | 57.9 |
| EGARCH(2,1) | 40.1 | 31.6 | 82.7 | 31.0 | 25.8 | 23.7 | 21.3 | 61.1 | 83.9 | 48.6 | 71.4 | 92.1 | 84.2 | 83.6 | 55.8 |
| EGARCH(1,2) | 80.5 | 55.3 | 31.6 | 35.6 | 47.1 | 59.9 | 49.8 | 62.6 | 86.3 | 44.4 | 62.6 | 97.0 | 94.2 | 93.9 | 64.4 |
| EGARCH(2,2) | 15.8 | 14.9 | 12.8 | 18.2 | 14.9 | 15.5 | 14.9 | 97.9 | 98.5 | 66.9 | 98.8 | 97.6 | 91.8 | 88.4 | 53.3 |
| NGARCH(1,1) | 51.4 | 53.8 | 94.2 | 62.9 | 53.2 | 32.8 | 36.8 | 93.3 | 90.9 | 83.0 | 93.6 | 54.4 | 53.5 | 47.7 | 64.4 |
| NGARCH(2,1) | 55.0 | 55.0 | 96.4 | 65.7 | 52.3 | 36.2 | 40.4 | 93.6 | 87.8 | 88.8 | 96.7 | 50.8 | 46.8 | 45.0 | 65.0 |
| NGARCH(1,2) | 58.7 | 55.9 | 77.2 | 55.3 | 55.0 | 42.2 | 42.6 | 95.1 | 93.9 | 85.1 | 96.0 | 57.8 | 48.9 | 45.9 | 65.0 |
| NGARCH(2,2) | 65.7 | 71.4 | 65.7 | 76.9 | 75.1 | 51.1 | 53.8 | 92.1 | 87.5 | 89.1 | 97.0 | 51.4 | 46.5 | 44.4 | 69.1 |
| A-PARCH(1,1) | 35.0 | 37.7 | 87.2 | 48.6 | 41.9 | 24.6 | 25.2 | 88.8 | 95.7 | 54.4 | 79.9 | 81.2 | 72.9 | 69.9 | 60.2 |
| A-PARCH(2,1) | 29.2 | 28.3 | 79.6 | 37.1 | 33.7 | 18.8 | 20.7 | 92.4 | 97.6 | 64.7 | 97.3 | 76.3 | 68.7 | 67.5 | 58.0 |
| A-PARCH(1,2) | 46.2 | 46.5 | 50.5 | 44.4 | 48.9 | 45.6 | 44.7 | 89.4 | 97.0 | 60.2 | 91.5 | 82.7 | 72.6 | 69.6 | 63.5 |
| A-PARCH(2,2) | 25.8 | 28.9 | 37.7 | 31.9 | 39.2 | 27.4 | 31.0 | **100.0** | **100.0** | 81.5 | **100.0** | 96.7 | 87.8 | 83.0 | 62.2 |
| GQ-ARCH(1,1) | 73.6 | 70.8 | 73.9 | 78.7 | 68.1 | 66.3 | 68.1 | 35.0 | 50.2 | 22.2 | 33.7 | 84.5 | 86.3 | 86.9 | 64.2 |
| GQ-ARCH(2,1) | 74.8 | 88.1 | 99.1 | 98.8 | 90.3 | 71.1 | 83.9 | 6.7 | 9.4 | 8.8 | 4.9 | 16.4 | 29.8 | 26.4 | 50.6 |
| GQ-ARCH(1,2) | 59.3 | 50.8 | 26.4 | 43.2 | 56.5 | 63.2 | 64.7 | 32.8 | 47.4 | 23.4 | 36.2 | 77.5 | 77.8 | 78.1 | 52.7 |
| GQ-ARCH(2,2) | 18.5 | 20.4 | 13.1 | 21.0 | 21.3 | 17.0 | 20.4 | 15.5 | 13.4 | 10.9 | 10.6 | 16.1 | 26.4 | 22.5 | 17.7 |
| H-GARCH(1,1) | 76.9 | 79.9 | 88.1 | 90.0 | 81.8 | 59.0 | 59.3 | 96.0 | 81.8 | 61.4 | 69.9 | 46.8 | 45.6 | 45.6 | 70.1 |
| AUG-GARCH(1,1) | 45.3 | 51.1 | 88.4 | 61.1 | 54.7 | 34.3 | 38.9 | 95.4 | 84.5 | 62.0 | 78.4 | 49.5 | 42.2 | 41.6 | 59.1 |

Relative performance ranking. Each row corresponds to a particular model, and a score shows the percentage of models (out of the total of 333) that performed worse than the particular model, measures 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),. . . , (11), are here denoted by $L_1, \ldots, L_7$. The last column is the average of the 14 scores.

Table 9: Models with $t$-distributed errors and constant mean

| Model | Exchange Rate Data | | | | | | | IBM Data | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | |
| ARCH(1) | 3.6 | 2.1 | 1.8 | .3 | 1.2 | 7.0 | 4.0 | .0 | .0 | .0 | .0 | 6.7 | 25.8 | 18.5 | 5.1 |
| GARCH(1,1) | 76.3 | 73.6 | 72.9 | 81.2 | 70.5 | 72.6 | 73.3 | 20.1 | 26.1 | 31.3 | 21.9 | 31.9 | 41.9 | 37.7 | 52.2 |
| GARCH(2,1) | 78.7 | 72.3 | 73.6 | 79.6 | 68.7 | 75.4 | 76.3 | 20.7 | 22.2 | 34.0 | 24.6 | 28.3 | 38.0 | 36.2 | 52.0 |
| GARCH(1,2) | 79.9 | 68.1 | 42.6 | 58.4 | 64.1 | 76.9 | 75.1 | 21.0 | 28.0 | 30.7 | 22.2 | 32.8 | 40.7 | 37.4 | 48.4 |
| GARCH(2,2) | 83.9 | 79.3 | 33.4 | 70.8 | 76.0 | 87.5 | 85.4 | 30.1 | 32.5 | 35.9 | 28.9 | 32.2 | 41.6 | 38.3 | 54.0 |
| IGARCH(1,1) | 1.5 | 1.2 | 54.1 | 4.6 | 2.4 | .9 | .9 | 10.0 | 2.1 | 94.5 | 12.2 | .3 | .3 | .3 | 13.2 |
| IGARCH(2,1) | 1.2 | 1.5 | 53.2 | 4.9 | 2.7 | 1.5 | 1.8 | 12.8 | 4.6 | 81.8 | 15.5 | 3.0 | 2.1 | 2.4 | 13.5 |
| IGARCH(1,2) | .9 | .9 | 51.1 | 4.0 | 2.1 | 1.2 | 1.2 | 10.9 | 3.0 | 92.1 | 13.4 | 1.2 | 1.2 | 1.2 | 13.2 |
| IGARCH(2,2) | 7.6 | 4.9 | 52.6 | 7.3 | 4.0 | 3.0 | 3.0 | 17.0 | 6.7 | 86.3 | 23.4 | 4.9 | 4.6 | 4.9 | 16.4 |
| TS-GARCH(1,1) | 34.3 | 31.0 | 93.3 | 38.3 | 32.2 | 16.4 | 16.7 | 78.1 | 80.9 | 68.1 | 83.3 | 56.8 | 50.5 | 46.5 | 51.9 |
| TS-GARCH(2,1) | 36.2 | 33.1 | 97.3 | 42.2 | 32.5 | 18.5 | 19.1 | 76.0 | 79.0 | 72.6 | 89.7 | 52.6 | 44.4 | 42.9 | 52.6 |
| TS-GARCH(1,2) | 50.5 | 44.1 | 81.5 | 45.9 | 41.3 | 31.0 | 29.8 | 78.7 | 83.0 | 68.7 | 88.8 | 59.3 | 47.1 | 44.7 | 56.7 |
| TS-GARCH(2,2) | 63.8 | 68.4 | 64.4 | 69.6 | 68.4 | 49.2 | 52.0 | 76.9 | 79.3 | 72.9 | 90.0 | 52.0 | 44.7 | 43.5 | 63.9 |
| A-GARCH(1,1) | 75.7 | 72.9 | 72.0 | 80.5 | 69.9 | 72.0 | 72.9 | 33.7 | 48.9 | 20.7 | 32.2 | 83.6 | 85.4 | 86.0 | 64.8 |
| A-GARCH(2,1) | 65.0 | 58.1 | 39.2 | 51.7 | 53.5 | 62.9 | 65.0 | 32.2 | 42.6 | 28.9 | 37.1 | 57.1 | 63.2 | 59.6 | 51.2 |
| A-GARCH(1,2) | 64.1 | 56.5 | 28.0 | 46.2 | 59.3 | 70.2 | 70.5 | 31.0 | 46.8 | 22.5 | 34.0 | 76.9 | 78.4 | 78.7 | 54.5 |
| A-GARCH(2,2) | 21.0 | 19.1 | 5.5 | 14.3 | 17.9 | 33.7 | 34.0 | 52.0 | 71.4 | 39.5 | 55.6 | 78.4 | 84.5 | 77.2 | 43.2 |
| NA-GARCH(1,1) | 72.6 | 71.1 | 70.2 | 77.8 | 67.8 | 68.7 | 69.0 | 39.8 | 57.1 | 16.7 | 38.3 | 94.8 | 90.9 | 92.7 | 66.3 |
| NA-GARCH(2,1) | 62.9 | 56.2 | 41.6 | 51.4 | 50.8 | 57.4 | 60.2 | 39.5 | 56.5 | 14.6 | 36.8 | 96.4 | 92.1 | 94.5 | 57.9 |
| NA-GARCH(1,2) | 63.5 | 57.1 | 25.8 | 45.0 | 61.4 | 73.3 | 74.2 | 38.0 | 54.7 | 14.9 | 39.8 | 93.3 | 88.8 | 89.7 | 58.5 |
| NA-GARCH(2,2) | 14.0 | 14.0 | 4.3 | 7.6 | 13.4 | 24.9 | 26.4 | 38.9 | 56.2 | 18.5 | 40.7 | 95.1 | 89.4 | 90.9 | 38.2 |
| V-GARCH(1,1) | 32.5 | 35.9 | 40.7 | 31.6 | 31.6 | 72.9 | 54.1 | 5.5 | 17.6 | 5.8 | 6.1 | 72.0 | 97.6 | 97.6 | 43.0 |
| V-GARCH(2,1) | 33.1 | 28.6 | 32.8 | 25.5 | 24.3 | 65.3 | 52.6 | 2.1 | 10.9 | 2.7 | 2.4 | 43.5 | 89.7 | 88.1 | 35.8 |
| V-GARCH(1,2) | 21.9 | 21.3 | 14.6 | 20.7 | 21.0 | 41.9 | 30.4 | 4.9 | 16.7 | 4.3 | 5.2 | 69.9 | 97.0 | 97.0 | 33.3 |
| V-GARCH(2,2) | 14.3 | 10.6 | 4.6 | 5.2 | 9.7 | 22.5 | 14.0 | 6.4 | 17.0 | 6.4 | 7.3 | 59.9 | 93.9 | 91.8 | 26.0 |
| THR-GARCH(1,1) | 25.5 | 24.3 | 84.5 | 28.9 | 23.7 | 13.1 | 13.1 | 74.5 | 89.1 | 50.8 | 73.6 | 90.3 | 76.0 | 75.4 | 53.0 |
| THR-GARCH(2,1) | 24.0 | 23.7 | 80.2 | 25.8 | 22.2 | 11.6 | 10.3 | 81.5 | 93.6 | 60.8 | 93.3 | 86.9 | 72.3 | 70.8 | 54.1 |
| THR-GARCH(1,2) | 45.9 | 43.2 | 68.1 | 42.6 | 42.6 | 39.2 | 39.5 | 73.9 | 90.0 | 52.6 | 79.3 | 91.5 | 76.9 | 76.0 | 61.5 |
| THR-GARCH(2,2) | 13.1 | 14.6 | 13.7 | 19.8 | 17.3 | 12.2 | 14.6 | 99.7 | 98.2 | 87.2 | 99.7 | 78.1 | 72.0 | 67.8 | 50.6 |
| GJR-GARCH(1,1) | 70.2 | 70.5 | 66.3 | 77.2 | 69.0 | 68.1 | 69.3 | 23.7 | 36.8 | 13.4 | 25.2 | 68.7 | 81.8 | 85.4 | 59.0 |
| GJR-GARCH(2,1) | 66.3 | 62.3 | 57.8 | 61.4 | 59.0 | 62.0 | 66.0 | 28.9 | 40.4 | 26.4 | 35.3 | 54.1 | 62.9 | 62.9 | 53.3 |
| GJR-GARCH(1,2) | 51.7 | 43.5 | 24.9 | 34.7 | 45.0 | 63.5 | 63.8 | 21.9 | 38.0 | 11.9 | 26.4 | 70.5 | 82.4 | 85.1 | 47.4 |
| GJR-GARCH(2,2) | 17.3 | 18.2 | 10.6 | 19.1 | 17.6 | 35.0 | 40.7 | 51.4 | 66.6 | 40.1 | 52.3 | 66.9 | 66.0 | 66.9 | 40.6 |
| LOG-GARCH(1,1) | 52.9 | 39.8 | 90.6 | 40.4 | 34.0 | 31.3 | 25.8 | 66.0 | 85.4 | 58.4 | 74.8 | 73.3 | 69.0 | 65.0 | 57.6 |
| LOG-GARCH(2,1) | 64.4 | 43.8 | 94.5 | 46.5 | 35.9 | 36.5 | 31.6 | 60.8 | 76.3 | 57.4 | 66.9 | 65.0 | 62.3 | 57.1 | 57.1 |
| LOG-GARCH(1,2) | 96.7 | 89.1 | 63.5 | 70.2 | 70.8 | 86.9 | 77.8 | 64.4 | 82.7 | 57.1 | 73.3 | 74.5 | 66.3 | 60.8 | 73.9 |
| LOG-GARCH(2,2) | 98.5 | 93.9 | 55.9 | 80.2 | 76.6 | 91.5 | 84.2 | 61.7 | 76.0 | 58.1 | 66.6 | 64.1 | 61.7 | 56.8 | 73.3 |
| EGARCH(1,1) | 37.7 | 31.9 | 83.0 | 32.8 | 26.7 | 25.2 | 23.4 | 64.1 | 90.3 | 43.5 | 61.4 | **100.0** | 96.0 | 96.0 | 58.0 |
| EGARCH(2,1) | 44.4 | 33.4 | 82.1 | 33.1 | 26.4 | 26.4 | 24.0 | 61.4 | 84.8 | 47.4 | 72.0 | 93.0 | 85.1 | 84.5 | 57.0 |
| EGARCH(1,2) | 68.1 | 45.9 | 25.2 | 29.2 | 43.5 | 58.4 | 48.6 | 62.3 | 86.6 | 43.8 | 62.3 | 97.9 | 94.5 | 94.2 | 61.5 |
| EGARCH(2,2) | 17.0 | 16.1 | 12.5 | 18.5 | 15.8 | 20.4 | 19.5 | 97.6 | 99.1 | 65.3 | 98.5 | 98.8 | 93.0 | 90.3 | 54.5 |
| NGARCH(1,1) | 51.1 | 52.0 | 90.3 | 59.6 | 48.3 | 34.0 | 38.0 | 95.7 | 93.3 | 79.6 | 94.5 | 57.4 | 57.4 | 49.5 | 64.4 |
| NGARCH(2,1) | 54.1 | 52.9 | 93.6 | 62.6 | 48.0 | 36.8 | 39.8 | 94.2 | 91.8 | 87.5 | 97.6 | 55.9 | 49.8 | 47.4 | 65.2 |
| NGARCH(1,2) | 62.0 | 57.8 | 79.0 | 56.2 | 52.9 | 43.2 | 42.9 | 96.4 | 95.1 | 82.7 | 96.4 | 63.5 | 54.4 | 48.6 | 66.5 |
| NGARCH(2,2) | 67.2 | 71.7 | 62.6 | 73.3 | 72.6 | 51.7 | 54.7 | 93.9 | 91.5 | 87.8 | 97.9 | 56.2 | 49.2 | 47.1 | 69.8 |
| A-PARCH(1,1) | 36.5 | 38.9 | 84.2 | 48.9 | 41.6 | 26.1 | 27.7 | 89.1 | 96.0 | 54.1 | 79.6 | 82.4 | 73.3 | 70.2 | 60.6 |
| A-PARCH(2,1) | 32.8 | 31.3 | 79.3 | 37.7 | 33.4 | 20.7 | 22.5 | 93.0 | 93.0 | 69.3 | 95.1 | 67.8 | 59.3 | 55.9 | 56.5 |
| A-PARCH(1,2) | 61.7 | 60.5 | 52.9 | 56.5 | 63.8 | 52.9 | 54.4 | 87.8 | 94.8 | 55.9 | 85.4 | 79.6 | 71.7 | 69.3 | 67.7 |
| A-PARCH(2,2) | 28.6 | 32.2 | 33.1 | 29.8 | 38.9 | 34.7 | 38.6 | 89.7 | 96.7 | 61.7 | 94.8 | 79.3 | 69.3 | 69.0 | 56.9 |
| GQ-ARCH(1,1) | 75.4 | 72.6 | 72.3 | 79.9 | 69.3 | 70.5 | 72.0 | 34.0 | 49.2 | 21.0 | 32.5 | 83.9 | 85.7 | 86.3 | 64.6 |
| GQ-ARCH(2,1) | 76.0 | 89.4 | 99.7 | 99.1 | 90.9 | 75.1 | 84.5 | 7.0 | 9.7 | 9.7 | 6.4 | 15.2 | 26.1 | 22.8 | 50.8 |
| GQ-ARCH(1,2) | 66.6 | 59.3 | 29.8 | 48.3 | 63.5 | 75.7 | 76.0 | 31.3 | 46.2 | 22.8 | 34.3 | 76.6 | 78.1 | 78.4 | 56.2 |
| GQ-ARCH(2,2) | 18.8 | 20.7 | 14.3 | 22.5 | 22.8 | 19.1 | 22.2 | 16.1 | 14.3 | 11.6 | 10.9 | 17.0 | 27.1 | 24.9 | 18.7 |
| H-GARCH(1,1) | 78.1 | 75.7 | 86.9 | 83.6 | 72.3 | 66.6 | 59.6 | 97.0 | 87.2 | 57.8 | 68.4 | 54.7 | 53.2 | 50.8 | 70.8 |
| AUG-GARCH(1,1) | 53.2 | 52.6 | 86.3 | 55.9 | 45.9 | 41.6 | 41.3 | 91.8 | 86.9 | 59.9 | 76.0 | 60.8 | 52.0 | 51.1 | 61.1 |

Relative performance ranking. Each row corresponds to a particular model, and a score shows the percentage of models (out of the total of 333) that performed worse than the particular model, measures 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),. . . , (11), are here denoted by $L_1, \ldots, L_7$. The last column is the average of the 14 scores.

Table 10: Models with $t$-distributed errors and GARCH-in-mean

| Model | Exchange Rate Data | | | | | | | IBM Data | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | |
| ARCH(1) | 4.3 | 2.7 | 2.4 | .6 | 1.8 | 7.3 | 4.3 | .6 | .6 | .3 | .3 | 7.0 | 25.2 | 17.3 | 5.3 |
| GARCH(1,1) | 59.9 | 63.5 | 76.0 | 70.5 | 62.9 | 54.4 | 57.1 | 22.2 | 31.3 | 30.4 | 22.8 | 34.3 | 46.2 | 42.2 | 48.1 |
| GARCH(2,1) | 64.7 | 66.0 | 76.3 | 72.9 | 64.4 | 58.7 | 61.1 | 23.1 | 26.7 | 33.1 | 25.5 | 31.6 | 40.4 | 38.0 | 48.8 |
| GARCH(1,2) | 69.9 | 61.4 | 43.5 | 54.4 | 58.7 | 62.6 | 62.9 | 22.8 | 32.2 | 30.1 | 24.3 | 34.7 | 45.3 | 41.3 | 46.0 |
| GARCH(2,2) | 91.5 | 98.8 | 53.5 | 98.5 | **100.0** | 97.3 | **100.0** | 30.7 | 33.4 | 35.6 | 29.5 | 33.7 | 43.8 | 40.4 | 63.3 |
| IGARCH(1,1) | .0 | .0 | 48.9 | 2.7 | .0 | .0 | .0 | 11.2 | 2.7 | 94.8 | 13.7 | .6 | .9 | .9 | 12.6 |
| IGARCH(2,1) | .6 | .6 | 50.8 | 3.3 | 1.5 | .6 | .6 | 13.4 | 5.2 | 80.9 | 18.2 | 4.3 | 2.7 | 3.3 | 13.3 |
| IGARCH(1,2) | .3 | .3 | 49.8 | 3.0 | .3 | .3 | .3 | 11.6 | 3.3 | 92.7 | 14.3 | 1.5 | 1.5 | 1.5 | 12.9 |
| IGARCH(2,2) | 5.5 | 2.4 | 43.2 | 3.6 | .9 | 2.7 | 1.5 | 17.6 | 7.6 | 86.0 | 24.9 | 5.5 | 5.2 | 5.2 | 15.1 |
| TS-GARCH(1,1) | 26.7 | 24.9 | 91.2 | 30.7 | 24.9 | 12.8 | 12.2 | 75.7 | 80.5 | 66.0 | 80.2 | 58.7 | 50.8 | 46.8 | 48.7 |
| TS-GARCH(2,1) | 28.9 | 26.1 | 95.4 | 35.3 | 27.4 | 13.7 | 14.3 | 75.4 | 79.9 | 70.8 | 88.4 | 53.5 | 45.0 | 43.8 | 49.8 |
| TS-GARCH(1,2) | 49.5 | 44.4 | 81.2 | 45.3 | 41.0 | 30.7 | 28.3 | 77.8 | 83.3 | 67.5 | 86.3 | 62.6 | 47.4 | 46.2 | 56.5 |
| TS-GARCH(2,2) | 55.3 | 65.7 | 61.4 | 64.4 | 67.2 | 47.1 | 49.2 | 76.3 | 79.6 | 71.7 | 89.1 | 52.9 | 44.1 | 43.2 | 61.9 |
| A-GARCH(1,1) | 62.3 | 66.3 | 78.4 | 75.1 | 65.3 | 55.3 | 58.4 | 35.9 | 51.1 | 19.5 | 32.8 | 86.3 | 87.2 | 87.2 | 61.5 |
| A-GARCH(2,1) | 52.6 | 53.2 | 46.5 | 52.0 | 50.2 | 52.6 | 55.6 | 35.3 | 44.1 | 28.3 | 38.0 | 58.1 | 63.5 | 59.9 | 49.3 |
| A-GARCH(1,2) | 57.8 | 49.5 | 27.7 | 39.5 | 52.6 | 60.8 | 61.4 | 33.4 | 48.6 | 20.4 | 35.6 | 79.0 | 80.9 | 81.8 | 52.1 |
| A-GARCH(2,2) | 19.1 | 18.8 | 6.7 | 13.7 | 16.4 | 28.3 | 28.9 | 52.6 | 72.9 | 39.2 | 55.9 | 80.2 | 86.6 | 81.2 | 42.9 |
| NA-GARCH(1,1) | 65.3 | 67.5 | 78.7 | 77.5 | 66.0 | 57.8 | 60.8 | 40.1 | 57.4 | 15.5 | 37.4 | 95.4 | 91.5 | 93.6 | 64.6 |
| NA-GARCH(2,1) | 53.5 | 54.4 | 47.7 | 52.6 | 50.5 | 52.3 | 55.3 | 41.0 | 58.4 | 17.3 | 39.2 | 94.5 | 90.3 | 92.1 | 57.1 |
| NA-GARCH(1,2) | 53.8 | 47.1 | 22.8 | 38.0 | 51.4 | 59.3 | 59.0 | 37.7 | 54.4 | 14.0 | 38.9 | 93.6 | 88.4 | 90.0 | 53.5 |
| NA-GARCH(2,2) | 12.2 | 12.2 | 2.7 | 6.1 | 11.2 | 20.1 | 22.8 | 38.6 | 55.3 | 16.4 | 40.1 | 94.2 | 89.1 | 90.6 | 36.5 |
| V-GARCH(1,1) | 31.0 | 35.0 | 45.9 | 32.2 | 31.3 | 67.8 | 53.2 | 5.8 | 17.9 | 5.5 | 5.8 | 72.9 | 98.2 | 98.2 | 42.9 |
| V-GARCH(2,1) | 33.4 | 30.4 | 35.0 | 26.4 | 24.6 | 64.4 | 52.3 | 2.4 | 11.6 | 2.4 | 2.1 | 47.1 | 90.0 | 89.4 | 36.5 |
| V-GARCH(1,2) | 22.8 | 22.5 | 15.2 | 21.3 | 21.6 | 45.9 | 34.7 | 2.7 | 11.9 | 1.8 | 2.7 | 62.0 | 96.4 | 96.7 | 32.7 |
| V-GARCH(2,2) | 14.6 | 10.3 | 5.8 | 5.5 | 9.4 | 21.3 | 12.8 | 12.5 | 31.0 | 9.1 | 10.3 | 70.2 | 95.1 | 95.1 | 28.8 |
| THR-GARCH(1,1) | 23.4 | 23.4 | 86.0 | 26.7 | 22.5 | 10.9 | 9.7 | 73.6 | 88.1 | 50.2 | 70.8 | 88.8 | 74.2 | 73.6 | 51.6 |
| THR-GARCH(2,1) | 22.5 | 21.6 | 80.5 | 22.8 | 20.1 | 9.1 | 9.1 | 80.9 | 91.2 | 61.1 | 93.0 | 79.9 | 68.1 | 68.7 | 52.0 |
| THR-GARCH(1,2) | 41.0 | 36.2 | 60.8 | 32.5 | 35.0 | 33.1 | 31.9 | 73.3 | 88.4 | 52.0 | 78.7 | 89.7 | 75.1 | 75.1 | 57.3 |
| THR-GARCH(2,2) | 10.9 | 12.8 | 12.2 | 17.3 | 12.5 | 9.4 | 10.0 | 99.1 | 99.4 | 76.3 | 99.1 | 91.8 | 79.6 | 74.8 | 50.4 |
| GJR-GARCH(1,1) | 57.4 | 62.6 | 72.6 | 69.9 | 63.2 | 54.1 | 56.8 | 24.3 | 37.1 | 14.3 | 25.8 | 69.3 | 82.1 | 85.7 | 55.4 |
| GJR-GARCH(2,1) | 52.3 | 56.8 | 61.1 | 59.3 | 57.4 | 53.5 | 57.4 | 29.8 | 40.1 | 26.7 | 35.9 | 52.3 | 61.4 | 58.1 | 50.2 |
| GJR-GARCH(1,2) | 41.9 | 35.3 | 21.9 | 27.4 | 36.5 | 55.0 | 53.5 | 23.4 | 38.9 | 12.5 | 27.1 | 70.8 | 81.2 | 84.8 | 43.6 |
| GJR-GARCH(2,2) | 16.4 | 16.4 | 10.3 | 17.6 | 14.3 | 32.2 | 37.1 | 51.7 | 66.9 | 40.7 | 53.8 | 66.3 | 65.7 | 66.6 | 39.7 |
| LOG-GARCH(1,1) | 38.9 | 32.8 | 89.7 | 34.3 | 28.9 | 24.3 | 21.0 | 65.7 | 85.1 | 56.8 | 71.7 | 75.1 | 69.6 | 66.0 | 54.3 |
| LOG-GARCH(2,1) | 46.8 | 37.4 | 93.9 | 38.6 | 31.0 | 28.0 | 24.9 | 59.0 | 75.4 | 56.2 | 65.7 | 66.0 | 61.1 | 56.5 | 52.9 |
| LOG-GARCH(1,2) | 95.7 | 79.6 | 66.6 | 59.0 | 60.2 | 81.5 | 68.4 | 63.5 | 82.1 | 55.3 | 71.1 | 75.7 | 66.9 | 62.6 | 70.6 |
| LOG-GARCH(2,2) | 96.0 | 90.9 | 58.7 | 76.0 | 73.9 | 90.3 | 80.2 | 59.3 | 75.7 | 56.5 | 66.0 | 65.3 | 60.8 | 56.2 | 71.8 |
| EGARCH(1,1) | 33.7 | 26.7 | 83.9 | 28.3 | 23.4 | 22.2 | 18.8 | 63.8 | 89.7 | 42.9 | 60.8 | 99.4 | 95.4 | 95.4 | 56.0 |
| EGARCH(2,1) | 35.3 | 27.4 | 81.8 | 28.0 | 23.1 | 22.8 | 20.1 | 60.2 | 84.2 | 47.1 | 70.5 | 92.4 | 83.9 | 83.3 | 54.3 |
| EGARCH(1,2) | 42.2 | 30.1 | 24.3 | 23.1 | 30.7 | 49.8 | 42.2 | 62.0 | 85.7 | 42.2 | 61.1 | 97.3 | 93.6 | 93.3 | 55.6 |
| EGARCH(2,2) | 13.4 | 11.9 | 11.9 | 13.4 | 10.9 | 13.4 | 10.6 | 98.2 | 98.8 | 65.7 | 98.2 | 98.5 | 92.4 | 88.8 | 51.8 |
| NGARCH(1,1) | 37.4 | 41.3 | 89.4 | 52.3 | 43.2 | 25.5 | 26.7 | 88.1 | 94.5 | 65.0 | 90.6 | 71.4 | 66.6 | 60.2 | 60.9 |
| NGARCH(2,1) | 39.8 | 42.9 | 92.7 | 55.0 | 44.4 | 26.7 | 29.5 | 90.9 | 92.4 | 79.0 | 95.7 | 62.3 | 54.1 | 49.8 | 61.1 |
| NGARCH(1,2) | 67.5 | 64.7 | 76.6 | 60.5 | 61.7 | 47.4 | 46.5 | 92.7 | 94.2 | 74.8 | 93.9 | 65.7 | 59.0 | 52.9 | 68.4 |
| NGARCH(2,2) | 54.7 | 65.3 | 61.7 | 63.5 | 66.9 | 46.8 | 48.3 | 90.3 | 92.1 | 78.4 | 95.4 | 61.7 | 54.7 | 50.5 | 66.5 |
| A-PARCH(1,1) | 32.2 | 35.6 | 85.7 | 46.8 | 39.8 | 23.4 | 24.6 | 87.5 | 95.4 | 52.3 | 78.1 | 85.4 | 77.2 | 73.9 | 59.9 |
| A-PARCH(2,1) | 26.1 | 25.5 | 80.9 | 34.0 | 29.8 | 14.6 | 16.4 | 86.0 | 97.9 | 45.0 | 63.8 | 98.2 | 92.7 | 91.2 | 57.3 |
| A-PARCH(1,2) | 39.2 | 33.7 | 47.4 | 28.6 | 32.8 | 35.9 | 33.1 | 81.8 | 97.3 | 41.6 | 60.2 | 99.7 | 96.7 | 96.4 | 58.9 |
| A-PARCH(2,2) | 27.4 | 26.4 | 31.9 | 26.1 | 31.9 | 28.9 | 28.0 | 90.0 | 92.7 | 62.3 | 92.1 | 69.6 | 63.8 | 59.3 | 52.2 |
| GQ-ARCH(1,1) | 62.6 | 66.6 | 78.1 | 75.4 | 65.7 | 55.9 | 58.7 | 36.2 | 51.4 | 19.8 | 33.1 | 86.6 | 87.5 | 87.5 | 61.8 |
| GQ-ARCH(2,1) | 50.2 | 74.8 | **100.0** | 98.2 | 83.6 | 53.2 | 76.9 | 7.6 | 10.6 | 9.4 | 7.0 | 16.7 | 28.3 | 25.8 | 45.9 |
| GQ-ARCH(1,2) | 59.0 | 50.5 | 26.7 | 39.8 | 53.8 | 61.4 | 62.0 | 32.5 | 48.3 | 20.1 | 35.0 | 78.7 | 80.5 | 81.5 | 52.1 |
| GQ-ARCH(2,2) | 14.9 | 16.7 | 14.0 | 20.1 | 18.2 | 12.5 | 13.7 | 17.9 | 16.1 | 15.8 | 11.6 | 18.8 | 32.2 | 30.1 | 18.0 |
| H-GARCH(1,1) | 55.6 | 67.8 | 90.0 | 85.7 | 74.5 | 45.0 | 47.4 | 90.6 | 96.4 | 49.8 | 67.2 | 82.1 | 79.3 | 74.5 | 71.8 |
| AUG-GARCH(1,1) | 29.5 | 41.9 | 92.4 | 61.7 | 57.8 | 23.1 | 26.1 | 91.2 | 86.0 | 60.5 | 75.7 | 59.0 | 49.5 | 49.2 | 57.4 |

Relative performance ranking. Each row corresponds to a particular model, and a score shows the percentage of models (out of the total of 333) that performed worse than the particular model, measures 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),. . . , (11), are here denoted by $L_1, \ldots, L_7$. The last column is the average of the 14 scores.

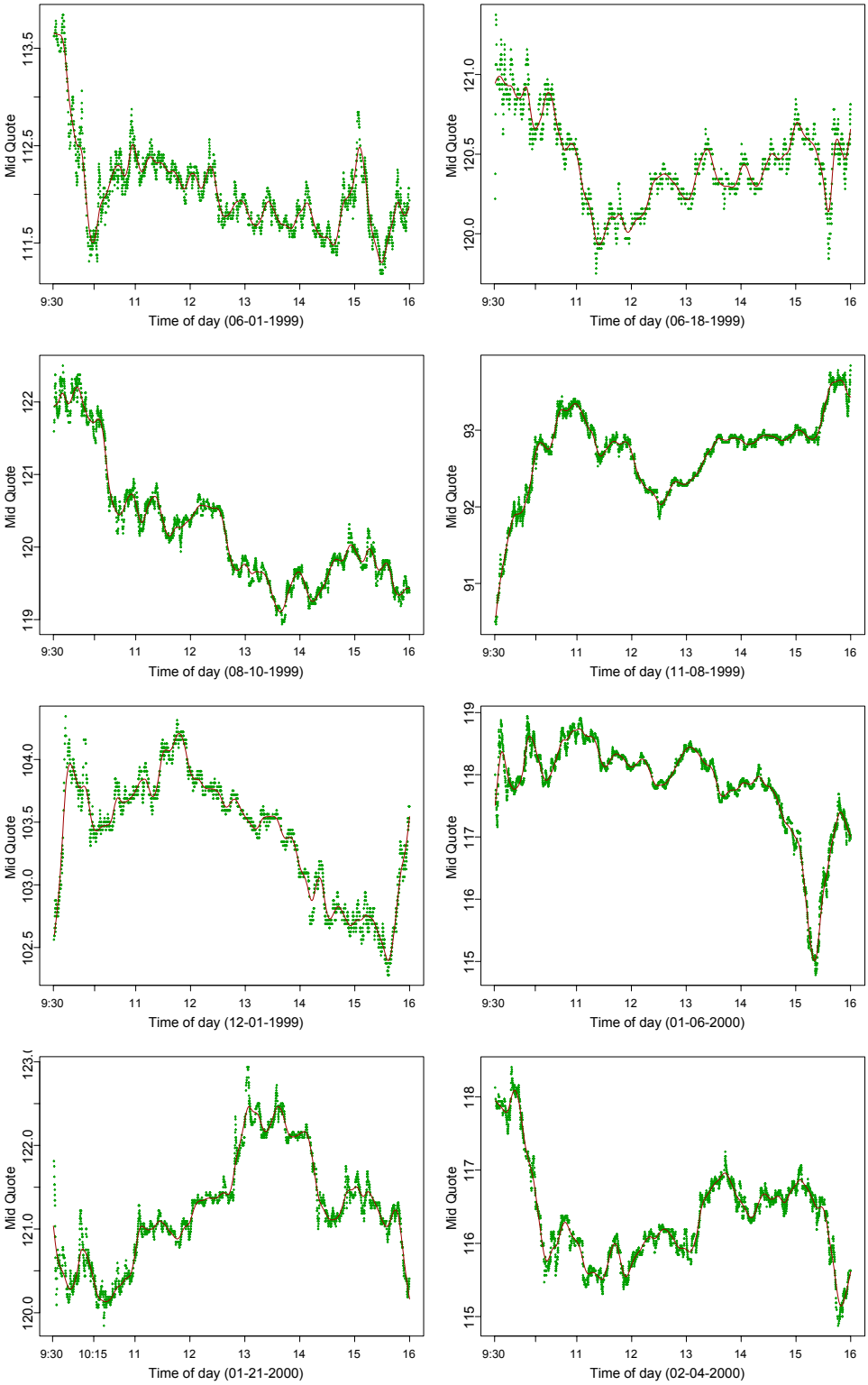


Figure 1: Intra-day mid quotes, and fitted spline-curves.

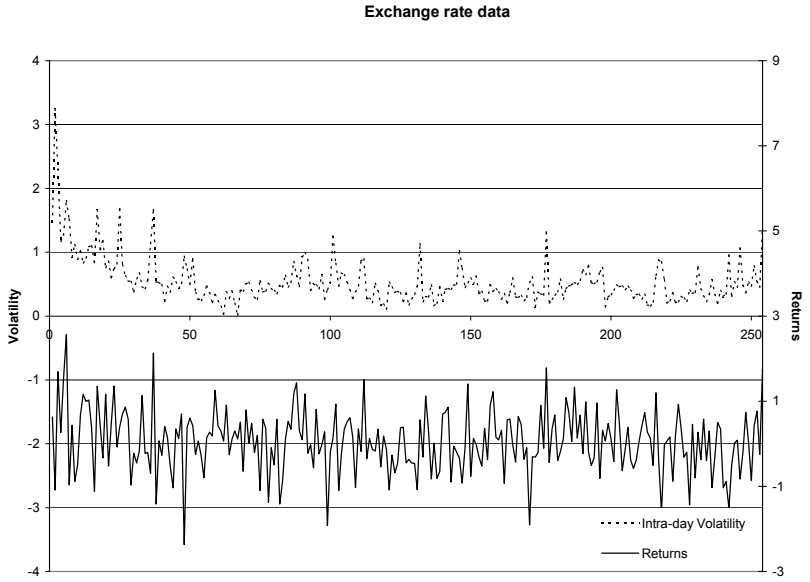


Figure 2: The intra-day volatility and returns of the DM-$ exchange rate data.

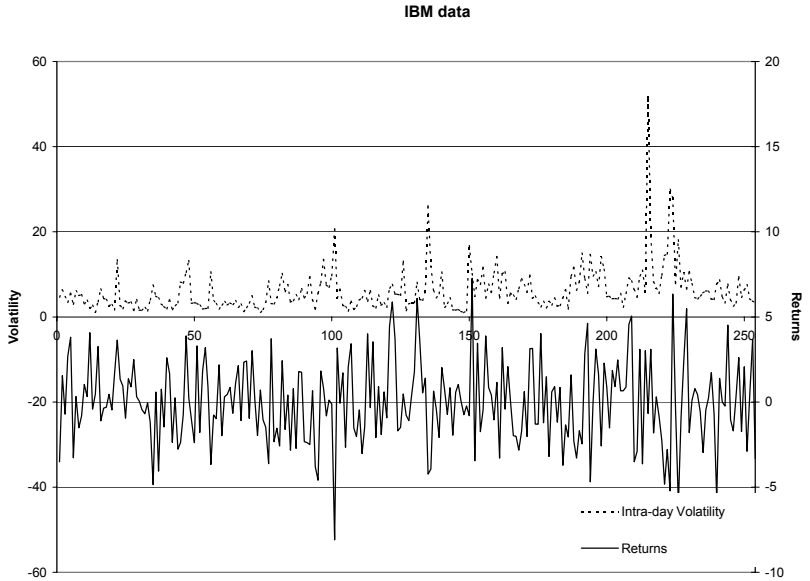


Figure 3: The intra-day volatility and returns of the DM-$ IBM data.