Hansen, Peter Reinhard; Lunde, Asger; Nason, James M.

Working Paper

Choosing the best volatility models: the model confidence set approach

Working Paper, No. 2003-28

Provided in Cooperation with:
Federal Reserve Bank of Atlanta

Suggested Citation: Hansen, Peter Reinhard; Lunde, Asger; Nason, James M. (2003) : Choosing the best volatility models: the model confidence set approach, Working Paper, No. 2003-28, Federal Reserve Bank of Atlanta, Atlanta, GA

This Version is available at:
https://hdl.handle.net/10419/100842

FEDERAL
RESERVE
BANK
of ATLANTA

Choosing the Best Volatility Models:
The Model Confidence Set Approach

Peter Reinhard Hansen, Asger Lunde, and James M. Nason

Working Paper 2003-28
October 2003

Working Paper Series

Federal Reserve Bank of Atlanta
Working Paper 2003-28
October 2003

Choosing the Best Volatility Models: The Model Confidence Set Approach

Peter Reinhard Hansen, Brown University
Asger Lunde, Aarhus University
James M. Nason, Federal Reserve Bank of Atlanta

Abstract: This paper applies the model confidence sets (MCS) procedure to a set of volatility models. A MSC is analogous to a confidence interval of parameter in the sense that the former contains the best forecasting model with a certain probability. The key to the MCS is that it acknowledges the limitations of the information in the data. The empirical exercise is based on fifty-five volatility models, and the MCS includes about a third of these when evaluated by mean square error, whereas the MCS contains only a VGARCH model when mean absolute deviation criterion is used. We conduct a simulation study that shows the MCS captures the superior models across a range of significance levels. When we benchmark the MCS relative to a Bonferroni bound, this bound delivers inferior performance.

JEL classification: C12, C19, C44, C52, C53

Key words: forecasting, model selection, multiple comparison, data mining

---

The authors thank Mark Kamstra and seminar pariticpants at the Federal Reserve Bank of Atlanta for useful comments. Financial support from the Danish Research Agency, grant no. 24-00-0363, and the Salomon Research Award at Brown University is gratefully acknowledged. This paper also owes much to the Federal Reserve Bank of Atlanta, which provided. support and hospitality to the first and third authors. The views expressed here are the authors' and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility.

Please address questions regarding content to Peter Reinhard Hansen, Department of Economics, Box B, Brown University, Providence, Rhode Island 02912, Peter_Hansen@brown.edu; Asger Lunde, The Aarhus School of Business, Department of Information Science, Fuglesangs Alle 4 DK-8210, Aarhus V, Denmark, alunde@asb.dk, or James M. Nason, Federal Reserve Bank of Atlanta, Research Department, 1000 Peachtree Street, N.E., Atlanta, Georgia 30309, 404-498-8891, jim.nason@atl.frb.org.

Choosing the Best Volatility Models: The Model Confidence Set Approach

I. Introduction

The literature on volatility models of asset returns has proposed a large number of specifications, starting with the ARCH model of Engle (1982) and the GARCH model of Bollerslev (1986). This gives practitioners a wide range of models to choose from and naturally leads to the question: *Which is the best volatility model?* It is difficult to answer this question because asset returns often do not contain sufficient information to identify a single volatility model as "best".

This paper offers some resolution of this quandary. We characterize the volatility models of stock returns that significantly dominate others, in an out-of-sample setting. The metric for assessing the

forecasts of volatility models is the Model Confidence Set (MCS) method of Hansen et al. (2003). The MCS is reviewed in this paper along with a discussion of the inferences that can be drawn from the MCS. We study the properties of the MCS through simulation experiments and find generally positive results. The MCS is also benchmarked to a Bonferroni bound approach, and we find the latter to be inferior to the former.

The MCS is an innovative approach to the problem of picking the "best" forecasting model, in this case volatility models, in an out-of-sample evaluation under a loss function specified by the user. The interpretation of a MCS is that of a confidence interval for a parameter, in the sense that a MCS contains the best model with a given level of confidence. Hence, the difference between the MCS and other selection criterion is analogous to the difference between the confidence interval of a parameter and a point estimate of a parameter.

The MCS procedure represents a general approach to model selection. This method neither assumes knowledge of the correct specification, nor does it require that the "true" model is available as one of the competing models. Another advantage is that a MCS does not discard a model unless it is found to be *significantly* inferior relative to other models. It is also more appealing to work with a set of forecasting models because in practice it often cannot be ruled out that two (or more) competing models are equally good (in population). This suggests the MCS dominates methods that require a single model be selected as "best".

The MCS accounts for uncertainty across a set of forecasting models, rather than uncertainty about the "best" model. The MCS method acknowledges the limitations of the information contained in the data by selecting a set of models, unlike more commonly used model selection criteria which select a single model. However, the MCS can consist solely of the best model. In this case, the MCS signals this model performs significantly better than all other models under consideration.

The framework in which one might test for superior predictive ability (SPA), is very similar to that of the MCS method. SPA tests can test whether a particular 'benchmark' model is significantly outperformed by other models, so a SPA test is suitable when the focal point is particular model. The MCS approach has three advantages over tests for SPA. First, the MCS procedure is 'benchmark free', because it does not require a benchmark model to be specified as is the case for SPA tests. Second, the MCS method characterizes the entire set of models that are/are not significantly out-performed by other models, while a test for SPA only provides evidence about the relative performance of a single model (the benchmark). Third, the MCS method relies on tests of simple hypotheses as opposed to composite hypotheses. This is important because this avoids the potential problems of SPA tests in which composite

hypotheses are examined, see Hansen (2003*a*). For example, the reality check of White (2000) which is a test for SPA, is sensitive to the inclusion of poor (and irrelevant) models in the set of competing models, which is related to the composite form of the null hypothesis. The test for SPA by Hansen (2001) avoids this problem by reducing the influence of poor performing models, and in the present framework the poor performing models are eliminated in the early stage of the MCS procedure, which also solves the problem.

This paper is concerned with the problem of drawing correct inference from multiple models. When many models are being compared a dimensionality problem arises, and this problem can be handled with bootstrap techniques. This approach was introduced by White (2000) to test for SPA, and further examined by Hansen (2001). This bootstrap approach was also used by Hansen (2003*b*) for in-sample analysis of two or more regression models. The advantage of the bootstrap method is that it circumvents the need for an explicit estimator of a high-dimensional covariance matrix. We apply the block-bootstrap in our analysis, where the bootstrap is used repeatedly in the sequential testing.

A MCS can be used to construct combination forecasts. The exact way that individual forecasts should be combined depends on several characteristics, including the loss function and the correlation structure of model forecasts. If the number of models in the MCS is too large to obtain a sensible estimate of the correlation structure a reasonable way to proceed is to weight the individual forecasts equally. The reason is that the MCS consists precisely of the models that cannot be rejected as being equally good. It is easy to analyze whether the combined forecast outperforms the individual forecasts, by deriving the MCS for the set of models, which includes the combined forecast as an additional model. If the combined forecast is superior, the MCS of the expanded set of forecasts should consist only of the combined forecast.

Construction of a MCS involves a sequence of tests for equal predictive ability (EPA). This trims the set of candidate models by deleting models that are found to be significantly inferior. The set of surviving models is the MCS, which is guaranteed to contain the best model with a certain level of confidence. The Pantula (1989) testing principle is used because it ensures that sequential testing does not distort the overall size of the test. Readers familiar with the trace-test for selecting the rank in the cointegrated vector autoregressive model will recognize this testing principle, see Johansen (1988, 1996).

Tests of EPA has been proposed by Diebold & Mariano (1995) and West (1996). Their tests are based on asymptotic normality of average (relative) performance, which require an estimate of the asymptotic covariance matrix. Since the dimension of the estimators of this covariance matrix increase with the number of models under consideration, the Diebold and Mariano and West EPA tests are useful for

comparing a moderate number of models, but become unwieldy when the objective is to compare a large number of models.

We implement the MCS with a study of volatility models of equity returns. Volatility models attempts to describe the intertemporal behavior of the volatility in equity returns. The literature contains a vast number of studies that evaluate and compare volatility models (see, e.g., Poon & Granger (2003) that contains a review of 93 papers). One example is Hansen & Lunde (2001) who report that a GARCH(1,1) model is not significantly outperformed in an analysis of exchange rates. On the other hand, they report the GARCH(1,1) model is inferior to other volatility models in an analysis of IBM stock returns. In their analysis, the best performing volatility models is primarily those that can accommodate a leverage effect.

We estimate and forecast volatility models of daily returns on the SPYDER[1] (Standard & Poor's Depository Receipts, Amex: SPY), which tracks the S&P 500 index. Daily stock returns taken from the Trade and Quote database on a sample that begins in January 3, 1995 and ends with February 28, 2002. The volatility models we examine are discussed in Hansen & Lunde (2001). Although it seems reasonable to expect that no volatility models of equity returns outperforms the rest, we report a surprisingly small MCS for one of the two criteria we use in the evaluation. When evaluating the models using the mean squared error, the MCS contains about a third of the 55 models. Evaluation of models using mean absolute deviation results in a MCS that includes only a VGARCH model, see Engle & Ng (1993). This model incorporates an asymmetry term that captures the leverage effect in a simple way, see Black (1976).

Evidence about the properties of the MCS is acquired through simulation experiments. These results are very encouraging because the MCS is shown to have the correct size and to become more powerful as the expected loss differential (between superior and inferior models) increases. We also compare the MCS to an alternative approach that employs the Bonferroni bound. These results clearly show that the Bonferroni bound method, in its simplest form, is not suitable for the problem of multiple comparison of models.

This paper is organized as follows. In Section 2, we describe the MCS method and its properties. In Section 3, we apply the MCS procedure to 55 volatility models using daily returns on the SPYDER, which tracks the S&P 500 index. Section 4 reports our simulation study of the MCS method and compares it to a related Bonferroni bound method. Section 5 summarizes.[2]

[1] See, e.g., Hasbrouck (2002) or Elton, Gruber, Comer & Li (2002)

[2] The two appendices of the paper contain assumptions and theoretical results for the MCS method and a detailed description of the bootstrap implementation.

II. Model Confidence Sets for Volatility Models

Let $\{p_t\}$ be a logarithmic price process where a unit of time corresponds to a trading day, such that the continuously compounded daily returns are given by $r_t = p_t - p_{t-1}$, for $t \in \mathbb{Z}$. We let $\mathcal{F}_t$ denote the information set at time t, and define the conditional mean, $\mu_t \equiv E(r_t|\mathcal{F}_{t-1})$, and the conditional variance, $\sigma_t^2 \equiv \text{var}(r_t|\mathcal{F}_{t-1})$. The sample is divided into an estimation period, $(t = -n_0, \ldots, 0)$, and an evaluation period, $(t = 1, \ldots, n)$. The evaluation is made conditional on the parameter estimates, in the sense that models are compared given these particular estimates.

We substitute a proxy for σ_t^2 because the conditional variance is unobserved. As has become common practice, we employ realized variance, $\hat{\sigma}_t^2$, which is a precise measure of σ_t^2.[3] In fact, it is important to use precise measures of the conditional variance to avoid an inconsistent ranking of the alternatives, see Hansen & Lunde (2003).

The objective is to determine which volatility model best describes the variation in σ_t^2. Table I lists the volatility models that we compare with loss functions that are evaluated out-of-sample. The models are indexed by $i = 1, \ldots, m$, and model i's forecasts of σ_t^2 is denoted by $h_{i,t}^2$. We rank models according to their expected loss using one of two loss functions: mean square error (MSE), $L(h_{i,t}^2, \sigma_t^2) = (h_{i,t}^2 - \sigma_t^2)^2$, and mean absolute deviation (MAD, $L(h_{i,t}^2, \sigma_t^2) = |h_{i,t}^2 - \sigma_t^2|$. The loss differential between models i and j, is given by,

$$d_{ij,t} = L(h_{i,t}^2, \hat{\sigma}_t^2) - L(h_{j,t}^2, \hat{\sigma}_t^2), \qquad i, j = 1, \ldots, m, \quad t = 1, \ldots, n.$$

We make the following assumption about the loss function differential.

Assumption 1 *$E\left|d_{ij,t}\right| < \infty$ and $\left\{d_{ij,t}\right\}$ is stationary and ergodic, for all $i, j = 1, \ldots, m$,*

It is important to note that Assumption 1 does not require $L(h_{i,t}^2, \hat{\sigma}_t^2)$ to be stationary and ergodic. Assumption 1 makes it possible to define the best volatility model(s) from a set of models under consideration that are denoted by $\mathcal{M}_0 = \{1, \ldots, m\}$.

Definition 1 (Set of superior models) *The set of superior models is defined by,*

$$\mathcal{M}^* \equiv \{i \in \mathcal{M}_0 : E(d_{ij,t}) \leq 0 \quad \textit{for all } j \in \mathcal{M}_0\}.$$

Our empirical problem is to determine which models are in $\mathcal{M}^*$ and which are not. A MCS set $\widehat{\mathcal{M}}_\alpha^*$ is guaranteed to contain any given model of $\mathcal{M}^*$ with a certain level of confidence, $1 - \alpha$, e.g., $1 - \alpha = 95\%$.

[3] Realized variance is an estimate of integrated variance, which, in turn, is an unbiased measure of the conditional variance, σ_t^2, see, Barndorff-Nielsen & Shephard (2002).

(i) Estimation of Model Confidence Sets

The MCS is determined after sequentially trimming the set of candidate models, $\mathcal{M}_0$. At each step, the hypothesis of EPA,

$$H_0 : E(d_{ij,t}) = 0, \qquad \text{for all } i, j \in \mathcal{M}, \tag{1}$$

is tested for a set of models $\mathcal{M} \subset \mathcal{M}_0$. The first test is for the full set of candidate models, $\mathcal{M} = \mathcal{M}_0$, and if H_0 is rejected, the worst performing model is eliminated from $\mathcal{M}$. This trimming is repeated until the first non-rejection occurs, and the set of surviving models is the model confidence set, $\widehat{\mathcal{M}}^*_\alpha$. By holding the significance level, α, fixed at each step of the MCS procedure, we construct a $(1-\alpha)$-confidence set, $\widehat{\mathcal{M}}^*_\alpha$, for the best models in $\mathcal{M}_0$.

We define the variables, $\bar{d}_{ij} \equiv \frac{1}{n}\sum_{t=1}^{n} d_{ij,t}$ and $\bar{d}_{i\cdot} \equiv \frac{1}{m-1}\sum_{j\in\mathcal{M}} \bar{d}_{ij}$ for $i, j \in \mathcal{M}$. The first variable, $\bar{d}_{ij}$, measures relative performance between model i and j, whereas $\bar{d}_{i\cdot}$ can be thought of as model i's performance relative to the average of the models in $\mathcal{M}$. The measure we use to rank the models in $\mathcal{M}$ is $v_{i,\mathcal{M}} \equiv \bar{d}_{i\cdot}/\sqrt{\widehat{\text{var}}(\bar{d}_{i\cdot})}$, where $\widehat{\text{var}}(\bar{d}_{i\cdot})$ is a bootstrap estimate of $\text{var}(\bar{d}_{i\cdot})$, see Appendix B. The standardized loss of model i relative to the other models in $\mathcal{M}$ is given by $v_{i,\mathcal{M}}$, and the worst performing model is defined by $i^\dagger \equiv \arg\max_{i\in\mathcal{M}} v_{i,\mathcal{M}}$.

Thus the algorithm for constructing the MCS is as follows:

1. Initiate the procedure by setting $\mathcal{M} = \mathcal{M}_0$.

2. Test for EPA of the models in $\mathcal{M}$.

   (a) If the EPA-hypothesis is "accepted". We set $\widehat{\mathcal{M}}^*_\alpha \equiv \mathcal{M}$, and report the $(1-\alpha)$-confidence set.

   (b) If the EPA-hypothesis is rejected.

      i. Then define

      $$\bar{d}_{i\cdot} \equiv \frac{1}{m}\sum_{j\in\mathcal{M}} \bar{d}_{ij},$$

      where m denotes the number of models in $\mathcal{M}$. This statistic, $\bar{d}_{i\cdot}$, measures the performance of model i, relative to the average across models.

      ii. Determine the "worst performing model" from $\mathcal{M}$, as defined by

      $$i^\dagger \equiv \arg\max_{i\in\mathcal{M}} \frac{\bar{d}_{i\cdot}}{\sqrt{\widehat{\text{var}}(\bar{d}_{i\cdot})}},$$

      where $\widehat{\text{var}}(\bar{d}_{i\cdot})$ is an estimate of $\text{var}(\bar{d}_{i\cdot})$.

      iii. Remove model $i^\dagger$ from $\mathcal{M}$, and repeat step 2.

(ii) Tests

Our tests for EPA employ the *range statistic*, T_R, and the *semi-quadratic statistic*, T_{SQ}, given by

$$T_R = \max_{i,j\in\mathcal{M}} \frac{|\bar{d}_{ij}|}{\sqrt{\widehat{\text{var}}(\bar{d}_{ij})}}, \qquad \text{and} \qquad T_{SQ} = \sum_{i<j} \frac{(\bar{d}_{ij})^2}{\widehat{\text{var}}(\bar{d}_{ij})},$$

where the sum is taken over the models in $\mathcal{M}$, and $\widehat{\text{var}}(\bar{d}_{ij})$ is an estimate of $\text{var}(\bar{d}_{ij})$ that we obtain from the bootstrap implementation.

We require an estimate of the (asymptotic) distributions of T_R and T_{SQ} to test for EPA under the null hypothesis of (1). The covariances, $\text{cov}(\bar{d}_{ij}, \bar{d}_{kl})$, $i, j, k, l \in \mathcal{M}$, are nuisance parameters in this context because the distributions of T_R and T_{SQ} depend on them. Fortunately the bootstrap is capable of estimating these distributions consistently, under mild regularity conditions. Alternatively, the asymptotic distributions can be bounded (over the nuisance parameters) using the Bonferroni bound. Although the appeal of the Bonferroni bound method is its simplicity, a drawback this method is that it can be very conservative, which results in very poor power properties.

The Bonferroni bound is based on the observation that $T_R = \max_{i,j\in\mathcal{M}} |t_{ij}|$, where $t_{ij} \equiv \bar{d}_{ij}/\sqrt{\widehat{\text{var}}(\bar{d}_{ij})}$ and the inequality (Bonferroni bound)

$$P(\max_{i,j\in\mathcal{M}} |t_{ij}| < x) \leq \sum_{\substack{i,j\in\mathcal{M} \\ i\neq j}} P(|t_{ij}| < x). \tag{2}$$

Under standard regularity conditions, it holds that $t_{ij} \overset{A}{\sim} N(0,1)$, such that the right hand side of (2) can be approximated by $[\Phi(x) - \Phi(-x)]\, m(m-1)/2$, where Φ is the cdf of the standard normal. If we apply a $\frac{2\alpha}{m(m-1)}$ critical value from the standard normal, the result is a conservative test with significance level α. This is the simplest form of the Bonferroni bound method, and as can be expected, the power of this method will decrease as m increases. We compare the properties of the Bonferroni bound method to those of the bootstrap method with simulation experiments.

We need to strengthen Assumption 1 to validate our bootstrap implementation. Let $\mathbf{d}_t$ denote a $m(m-1)/2$-dimensional vector whose elements are given by $d_{ij,t}$ for $i < j$, and $i, j \in \mathcal{M}$.

Assumption 2 *The process,* $\{\mathbf{d}_t\}$*, is geometric strong mixing and* $E\,|\mathbf{d}_t|^{2+\delta} < \infty$ *for some* $\delta > 0$.[4]

The sample average, $\bar{\mathbf{d}} = n^{-1}\sum_{t=1}^{n} \mathbf{d}_t$, is asymptotically normally distributed under Assumptions 1 and 2. The limit distributions of the test statistics we consider below, are given from this distribution.

The bootstrap avoids the need to obtain an estimate of the (asymptotic) covariance matrix of $\bar{\mathbf{d}}$. Since $\bar{\mathbf{d}}$ has dimension $m(m-1)/2$, it will often be very challenging, or impossible, to work with the limit

[4] See Härdle, Horowitz & Kreiss (2002) for details and references.

distribution of $\bar{\mathbf{d}}$. Rather than attempt to estimate the limit distribution of $\bar{\mathbf{d}}$ and transform it into the limit distribution of our test statistics, our bootstrap procedure operates on the distribution of a MCS statistic. The result is a univariate distribution which avoids the curse of dimensionality.

(iii) MCS p-values

A very useful feature of the MCS procedure is that it yields p-values for all models under consideration. The MCS p-value has the interpretation that a model with a small p-value is unlikely to be a member of the set of superior models, $\mathcal{M}^*$. We denote the p-value for model i as $\hat{p}_i$, $i = 1, \ldots, m$, such that

$$i \in \widehat{\mathcal{M}}^*_\alpha \text{ for } \alpha \leq \hat{p}_i \qquad \text{and} \qquad i \notin \widehat{\mathcal{M}}^*_\alpha \text{ for } \alpha > \hat{p}_i. \tag{3}$$

A MCS p-value is constructed as follows. Index the possible $m-1$ tests for EPA, by $k = 1, \ldots m-1$, and let $p(k)$ be the p-value of the kth test.[5] Since H_0 is rejected for large values of the test statistic, the p-value, $p(k)$, is given by $1 - F_k(t_k)$, where F_k is the (estimated) cdf of the test statistic and t_k is the observed test statistic of the kth test. We compute the test, $\kappa(i)$, that would have resulted in the model i being deleted from $\mathcal{M}$ for each model $i = 1, \ldots, m$. If model i is the model that survives all $m-1$ tests (is the best performing model, i^*), we use the convention, $\kappa(i^*) = m$, and $p(m) = 1$. The p-value of model i is defined by

$$\hat{p}_i = \max_{k \leq \kappa(i)} p(k).$$

If $\alpha < \hat{p}_i$, there is a test that results in non-rejection prior to the elimination of model i. Conversely, model i is when $\alpha \geq \hat{p}_i$.

III. Empirical Analysis

The conditional variance, σ_t^2, is our main object of interest. We examine 55 specifications for σ_t^2 that make up our set of candidate models, $\mathcal{M}_0$. For all models, we use a constant specification for the conditional mean, $\mu_t = \mu$, and take the standardized innovations, $e_t \equiv (r_t - \mu)/\sigma_t$, to be Gaussian.[6] The parametric specifications for σ_t include the GARCH model of Bollerslev (1986), the IGARCH model, the Taylor (1986)/Schwert (1989) TS-GARCH model, the A-GARCH, the NA-GARCH and the V-GARCH models suggested by Engle & Ng (1993), the threshold GARCH model (Thr.-GARCH) of Zakoian (1994), the GJR-GARCH model of Glosten, Jagannathan & Runkle (1993), the log-ARCH developed by Geweke (1986) and Pantula (1986), the EGARCH of Nelson (1991), the A-PARCH model

[5] Naturally, to compute the MCS does not require all possible tests to be conducted, unless $\widehat{\mathcal{M}}^*_\alpha$ consists of a single model.

[6] Unreported results for a larger set of models that also includes a t-distributed specification led to almost identical results for the predictive ability. The reason is that the volatility forecast of each of the models did not differ substantially for the two distributional specification.

proposed in Ding, Granger & Engle (1993), which nests the NGARCH of Higgins & Bera (1992)), and the GQ-ARCH proposed by Sentana (1995). Many of these models are nested within the general specifications studied in Hentshel (1995) and Duan (1997).[7] We also include the FIGARCH and FIEGARCH suggested by Baillie, Bollerslev & Mikkelsen (1996), and finally the canonical stochastic volatility model. All models are estimated by maximum likelihood.[8] Table I summarizes the volatility specifications.

[TABLE I]

It is worthwhile to note that some specifications are expressed in terms of the innovation, $\varepsilon_t \equiv (r_t - \mu)$, while others are expressed in terms of the standardized innovation, $e_t \equiv \varepsilon_t/\sigma_t$. The specifications have two lag length parameters, denoted p and q. We estimate the models over $p, q = 1, 2$, with the exceptions of the FIGARCH and FIEGARCH models where $p, q = 0, 1$.

(i) Data

The data are SPY transaction prices from the Trade and Quote (TAQ) database. SPY includes all trades and quotes from the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and National Association of Securities Dealers Automated Quotation (Nasdaq) securities. Our measure of realized variance is based on transactions from 9:30am to 4pm on a sample period that runs from January 3, 1995 to February 28, 2002, $T = 1802$ trading days. Prior to estimation the following observations are removed from the sample: *(i)* prices that are equal to zero, *(ii)* transactions that occurred outside the 9:30am to 4pm trading day, *(iii)* prices that deviate more than a 1% deviation from a local mean (of 100 transactions), *(iv)* prices that changed by 0.5% or more but 'changed back' within the next ten trades. This reduced the number of intra-day observations from 2,991,154 to 2,886,650, but retained 1802 close-to-close returns.

(ii) Empirical Results

The parameters of the volatility models are estimated using the first 1000 inter-day observations. These parameter estimates are used to make one-step-ahead forecasts for the following 802 periods. During the evaluation period, we estimate a measure for daily volatility (realized variance) using intra-day transac-

[7] Loudon, Watt & Yadav (2000) compare several models that are nested in the Aug-GARCH model.

[8] The likelihoods were maximized using Ox procedures. The simplex method was used for all models, except the stochastic volatility model, which was estimated with a modified version of `SVModel.ox`, see Koopman, Shephard & J.A.Doornik (1999). This code is found at `http://staff.feweb.vu.nl/koopman/sv/`.

tions and the Fourier method[9]; see Malliavin & Mancino (2002), Barucci & Reno (2002), and Hansen & Lunde (2003). The MCS are presented in Table II.

[TABLE II]

For the semi-quadratic statistic the MCS is presented in Figure 1 as a MSE/MAD scatter-plot.

[FIGURE 1]

Table II and Figure 1 present the MCS results for the volatility models under consideration. The T_{SQ} test results in a MCS that contains eight models with a confidence level of 75% and 19 models with 90% confidence under the MSE loss function. Using the same loss function, the more conservative Range-test, T_R, lands 23 models in the MCS given either confidence levels. The two tests leads to the same inference when the MAD loss function is used, in which case the MCS contains only a single model, the VGARCH12 model. The fact that the MAD criterion leads to fewer models is not surprising because this criterion is less sensitive to outliers (large mispredictions), such that the data can be more informative about the relative performance using this criterion.

For both criteria we see that the best performing models have a *leverage effect*, which is the asymmetric response in volatility to positive and negative shocks. The leverage effect was first noted by Black (1976), who reported a negative correlation between stock returns and changes in their volatility. The leverage effect is important for understanding stock return dynamics, as it is observed empirically that the volatility tends to rise in response to bad news (defined as lower than expected returns) and tends to fall after good news.

Our empirical results reveal another interesting observation. Many of the best performing models have specifications for the variance that are based on the standardized innovation, e_t. In particular, the best performing model, the VGARCH12, is formulated in terms of the standardized innovations.

It may seem puzzling that the NAGARCH22 is in the MCS for MSE because its sample performance is poor, see model 21 in Table II and Figure 1. The explanation is found in Table III.

[TABLE III]

[9] An advantage of the Fourier method for constructing realized variance is that: All intra-day prices are used (as oppose to using artificial equi-spaced intra-day returns; and by discarding the Fourier terms that are associated with the highest frequencies, it provides a natural way to remove the potential bias from market micro effects. For more details, see the references cited in the text.

A poor performing model may still end up in the MCS if $\widehat{\text{var}}(\bar{d}_{i\cdot})$ is large. From the identity $\bar{d}_{i\cdot} = \bar{L}_{i,\cdot} - \bar{L}_{\cdot,\cdot}$, where $\bar{L}_{i,\cdot} \equiv n^{-1}\sum_{t=1}^{n} L_{i,t}$, $L_{\cdot,t} \equiv m^{-1}\sum_{j\in\mathcal{M}} L_{j,t}$, and $\bar{L}_{\cdot,\cdot} \equiv n^{-1}\sum_{t=1}^{n} L_{\cdot,t}$, we see that

$$\begin{aligned}\text{var}(\bar{d}_{i\cdot}) &= \text{var}(\bar{L}_{\cdot,\cdot}) + \text{var}(\bar{L}_{i,\cdot}) - 2\text{cov}(\bar{L}_{i,\cdot}, \bar{L}_{\cdot,\cdot}) \\ &= \text{var}(\bar{L}_{\cdot,\cdot})\left(1 + \frac{\text{var}(\bar{L}_{i,\cdot})}{\text{var}(\bar{L}_{\cdot,\cdot})} - 2\sqrt{\frac{\text{var}(\bar{L}_{i,\cdot})}{\text{var}(\bar{L}_{\cdot,\cdot})}}\text{corr}(\bar{L}_{i,\cdot}, \bar{L}_{\cdot,\cdot})\right),\end{aligned}$$

provided that the variances are well defined. So the relative size of $\text{var}(\bar{d}_{i\cdot})$ is determined by two factors. One factor is the ratio of the variance of model i's loss relative to the average loss, and the other factor is the correlation between model i's loss and the average loss. The intuition of this decomposition is that a high relative variance makes it difficult to determine whether model i is truly inferior, or was simply 'unlucky' in this sample. Similarly, if the correlation is small $\bar{L}_{i,\cdot}$ is not expected to be close to $\bar{L}_{\cdot,\cdot}$.

In Table III we report the decomposition of the variance, in terms of the relative variance and the correlation. The calculations are sample averages. This table shows that the NAGARCH22 has a relatively small correlation with the average performance of the other models.[10] Thus, the NGARCH22 ends up in the MCS using the MSE loss function because the data does not contain sufficient information to be able to infer that this model is inferior within the MCS.

IV. A Small Simulation Study of the MCS Method

We construct a simulation experiment to compare our bootstrap method Bonferroni bound method. In each simulation we evaluate how the models were classified, as displayed in Figure 2.

[FIGURE 2]

We repeat the simulation experiments H times, which yields $A_{11,h}$, $A_{12,h}$, $A_{21,h}$, and $A_{22,h}$ for $h = 1, \ldots, H$. From these variables, we compute summary statistics: the frequency that all superior models are contained in the MCS,

$$Q_{\mathcal{M}^*\subset\widehat{\mathcal{M}}^*_\alpha} = \hat{P}\{\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_\alpha\} = H^{-1}\sum\nolimits_{h=1}^{H} 1_{\{A_{21,h}=0\}},$$

where $1_{\{\cdot\}}$ is the indicator function; the frequency that the MCS equal was exactly equal to the true set of superior models,

$$Q_{\widehat{\mathcal{M}}^*_\alpha=\mathcal{M}^*} = \hat{P}\{\widehat{\mathcal{M}}^*_\alpha = \mathcal{M}^*\} = H^{-1}\sum\nolimits_{h=1}^{H} 1_{\{A_{12,h}=0\}}1_{\{A_{21,h}=0\}}.$$

[10] The correlation for the NAGARCH22 is 0.964. Most other models have a correlation in excess of 0.99.

We also define the proportion of the models in $\widehat{\mathcal{M}}^*_\alpha$ that are truly superior,

$$Q_{\mathcal{M}^*|\widehat{\mathcal{M}}^*_\alpha} = \hat{P}\{\mathcal{M}^*|\widehat{\mathcal{M}}^*_\alpha\} = H^{-1}\sum_{h=1}^{H} A_{11,h}/(A_{11,h}+A_{21,h}),$$

and the proportion of models classified as inferior that are truly inferior,

$$Q_{\complement\mathcal{M}^*|\complement\widehat{\mathcal{M}}^*_\alpha} = \hat{P}\{\complement\mathcal{M}^*|\complement\widehat{\mathcal{M}}^*_\alpha\} = H^{-1}\sum_{h=1}^{H} A_{22,h}/(A_{21,h}+A_{22,h}),$$

which represents the "power" of the MCS.

Our bootstrap method and the Bonferroni bound method requires an estimate of var($\bar{d}_{ij}$), to implement the tests. The estimate is easily obtained from the bootstrap resamples, however for the sake of comparability, we have use the true variance in the simulations reported in this paper. Unreported simulation experiments show that the properties of the MCS procedure are similar with an estimated is used.

[TABLE IV]

Table IV, A ($\alpha = .05$) and B ($\alpha = 0.10$) contain a simulation experiment where we have generated random variables, $L_{i,t} \sim$ iid $N(\mu_i, 1)$, $i = 1, \ldots, m$ and $t = 1, \ldots, n$, where

$$\mu_i = \begin{cases} 0 & \text{for} \quad i = 1, \ldots, m^* \\ \lambda/\sqrt{n} & \text{for} \quad i = m^*+1, \ldots, m, \end{cases}$$

for $\lambda = 1, 3, 5$, $\rho \in (0.20, .50, .80)$ where $\rho \equiv m^*/m$, and for $m = 10, 40$, and $n = 250$. $B = 1{,}000$. $n_{iter} = 4{,}000$. $\alpha = 0.05, 0.10$. The basic properties of the MCS procedure can be studied in this simple design. However, as pointed out by referees, one would expect time dependence in $L_{i,t}$ and non-Gaussian distributions in many applications, and the analysis of these important aspects for future research.

The table reports four summery statistics for the two bootstrap tests, (T_R and T_{SQ}), and the Bonferroni bound test. As can be seen, the Bonferroni bound test perform very poorly, as it is hardly capable of eliminating any of the inferior models. The problem is that the conservative Bonferroni test lacks power so λ needs to be very large before the Bonferroni method is able to tell good and bad models apart.[11] The properties of the MCS procedure are as could be expected. The frequency that any of the best models are contained in the MCS ($Q_{\widehat{\mathcal{M}}^*_\alpha|\mathcal{M}^*}$) is always greater than $(1-\alpha)$, and the MCS becomes better at separating the inferior models from the superior models, as the performance gap, λ, increases. The requirement that $\widehat{\mathcal{M}}^*_\alpha = \mathcal{M}^*$ is very demanding, yet as λ increases the equality occurs quite frequently. The range test is less powerful than is the semi-quadratic test, and will on average have more models

[11] In lack of space we only present simulation results for $\rho = 0.5$, as the same picture emerged for $\rho = 0.2$ and $\rho = 0.8$.

in the MCS. This is not surprising, because a property of the range test is that all superior models will be contained in the MCS, with a frequency no less than $1-\alpha$, see Hansen et al. (2003). This property is also confirmed by the relevant statistic: $Q_{\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_\alpha}$. The two tests, T_R and T_{SQ}, perform well in the simulations, but exhibit different strengths and weaknesses, so there is no evidence that one dominate the other.

V. Summary

This paper studies the model confidence set (MCS) method of Hansen et al. (2003). The MCS provides a useful tool to select the superior forecasting models from a larger set. The model whose forecasts produce the minimum expected loss defines the best model, where the loss function is selected by the user. The MCS contains the best model with a level of confidence, which is specified by the user. Generally, the number of models in the MCS will increase as the level of confidence increases. Thus, the set of models in a MCS can be interpreted in the same way that a confidence interval covers the part of the real line in which the true value of a parameter resides with a certain significance level. When the MCS (with a high level of confidence) contains only one model, this model is very likely to be the truly superior model, in terms of the loss function that was applied in the construction of the MCS.

The MCS does not require as much prior information as other methods for model selection. There is no need to assume knowledge of the "true" data generating process or the optimal forecasting model in order to apply the MCS. This focuses uncertainty on the set of forecasting models under consideration. Thus, the MCS recognizes that the limitations of the data imply that in most cases it is not possible to settle on a single model. Instead, the MCS usually chooses the models that perform significantly better than the rest.

We apply the MCS to a set of volatility models to judge their ability to forecast the volatility of stock returns. The MCS selects VGARCH over the other volatility models we examine when the evaluation is made with the mean absolute deviation loss function, and the MCS contains about a third of the candidate models when the models are evaluated using the mean square error loss function. Thus, the MCS is able to separate superior from inferior volatility models under these loss functions, even at a ten percent significance level. Naturally, our empirical results are specific to our set of candidate models (and sample period) and the MCS could have been different had we included additional models.

We study the properties of the MCS with a small simulation experiments. These experiments reveal that the MCS possess the ability to separate superior from inferior models. In particular, the simulations show that the MCS yields a set of models that contains all, or almost all, truly superior models, but

rarely *exactly* captures the true set of superior models, unless the difference in expected performance of superior and inferior models is large. Perhaps most importantly, the MCS faces few problems when it has to classify inferior models as truly inferior. Thus, the MCS represents a powerful tool to choose the best set of forecasting models.

Appendix A: Assumptions and Theoretical Results

Lemma 1 *Let* $\bar{\mathbf{d}} = n^{-1}\sum_{t=1}^{n}\mathbf{d}_t$. *Under Assumptions 1 and 2 it holds that*

$$n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\delta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}), \qquad \text{as } n \to \infty, \tag{A.1}$$

where $\boldsymbol{\delta} \equiv E(\mathbf{d}_t)$, *and* $\boldsymbol{\Omega}$ *is the asymptotic covariance matrix.*

Proof. Follows by the multivariate central limit theorem for stationary and ergodic processes. ■

We note that the limit distribution need not be Gaussian, when the evaluation of models is made unconditionally on parameter estimates. Situations where different limit distributions arise, are discussed in West & McCracken (1998) and Clark & McCracken (2001). The former studies the situation with various estimation schemes and the latter studies the case with nested models. However, the approach West & McCracken (1998) and Clark & McCracken (2001) develop violates our Assumption 1.

Theorem 2 *Under Assumptions 1 and 2, the asymptotic distribution of* T_R *and* T_{SQ}, *under the hypothesis* $E(d_{ij,t}) = 0$ *for all* $i, j \in \mathcal{M}$, *are given by a continuous transformation of* $N(0, \boldsymbol{\Omega})$.

Proof. Define $\gamma_{ij} = \text{var}(\sqrt{n}\bar{d}_{ij})$. Under H_0, we note that $\max_{i,j\in\mathcal{M}} \left|\sqrt{n}\bar{d}_{ij}\right| / \gamma_{ij}$ and $n\sum_{i<j}\bar{d}_{ij}^2/\gamma_{ij}^2$, are continuous function of the elements of $\bar{\mathbf{d}}$, for which $E(d_{ij,t}) = 0$. By the continuous mapping theorem it follows that the asymptotic distributions of T_R and T_{SQ} are given from the limit distribution of $\bar{\mathbf{d}}$, under the assumption that $\widehat{\text{var}}(\sqrt{n}\bar{d}_{ij})$ is consistent for $\text{var}(\sqrt{n}\bar{d}_{ij})$. ■

Appendix B: Bootstrap Implementation

The bootstrap implementation consists of three steps. In the first step, resamples are generated. The second step uses the resamples to estimate certain parameters, and the distribution of the test statistic of interest is estimated in step three.

1. (a) Evaluate the dependence in $d_{ij,t}$ to get the bootstrap dependence parameter (the block-length). One way to do this is to estimate an autoregressive process for $d_{ij,t}$ $i, j \in \mathcal{M}$, and determine

its lag-length, e.g., by the use of information criteria. The bootstrap block-length can be set to be the largest lag-length (over $i,\ j \in \mathcal{M}$).

(b) Generate B bootstrap resamples of $\{1, \ldots, n\}$, using the block bootstrap, using k as the block-length. This can be done as follows, where Υ denotes the uniform distribution over the integers $\{1, \ldots, n\}$.

i. Determine the number of blocks needed for a sample of size n. I.e., let v the smallest integer that satisfies $v \geq \frac{n}{k}$.

ii. For $a = 0, \ldots, v-1$, let $\tau_{b_{ak+1}} \sim iid\ \Upsilon$. (First element of each block.)

iii. Set $\tau_{ak+t} = \tau_{b_{ak+t-1}} + 1 \bmod n$, for $t = 2, \ldots, k$, and for $a = 0, \ldots, v-1$.

iv. The bth resample index is given by $(\tau_{b_1}, \ldots, \tau_{b_n})$.

v. This is repeated for all resamples $b = 1, \ldots, B$.

2. Estimate variance parameters as follows.

(a) Define the bootstrap re-samples averages, $\bar{d}^*_{b,ij} = \frac{1}{n}\sum_{t=1}^{n} d_{ij,\tau_{b_t}}$, and let the bootstrap estimated of $\text{var}(\bar{d}_{ij})$ be given by

$$\widehat{\text{var}}(\bar{d}_{ij}) \equiv \frac{1}{B}\sum_{b=1}^{B}(\bar{d}^*_{b,ij} - \bar{d}_{ij})^2, \qquad \text{for } i,\ j \in \mathcal{M}.$$

(b) Similarly, calculate $\bar{d}^*_{b,i,\cdot} \equiv \frac{1}{m-1}\sum_{j\in\mathcal{M}} \bar{d}^*_{b,ij}$, and the bootstrap estimate of $\text{var}(\bar{d}_{i\cdot})$, which is defined by

$$\widehat{\text{var}}(\bar{d}_{i\cdot}) \equiv \frac{1}{B}\sum_{b=1}^{B}(\bar{d}^*_{b,i\cdot} - \bar{d}_{i\cdot})^2, \qquad \text{for } i \in \mathcal{M}.$$

3. Calculate the test statistic of interest, and obtain its bootstrap distribution.

(a) Calculate the two statistics, T_R and T_{SQ}.

(b) The estimated bootstrap distributions of T_R and T_{SQ}, under the null hypothesis, are given by the empirical distribution of

$$T^*_{b,R} = \max_{i,j\in\mathcal{M}'} \frac{\left|\bar{d}^*_{b,ij} - \bar{d}_{ij}\right|}{\sqrt{\widehat{\text{var}}(\bar{d}_{ij})}} \qquad \text{and} \qquad T^*_{b,SQ} = \sum_{\substack{i,j\in\mathcal{M}\\ i<j}} \frac{(\bar{d}^*_{b,ij} - \bar{d}_{ij})^2}{\widehat{\text{var}}(\bar{d}_{ij})},$$

respectively.

(c) A p-value of the test for EPA is obtained by calculating the ratio of $T^*_{b,R}$ (or $T^*_{b,SQ}$) that exceed T_R (or T_{SQ}). I.e.,

$$\hat{p}_R \equiv \frac{1}{B}\sum_{b=1}^{B} 1\left\{T^*_{b,R} > T_R\right\}, \qquad \text{and} \qquad \hat{p}_{SQ} \equiv \frac{1}{B}\sum_{b=1}^{B} 1\left\{T^*_{b,SQ} > T_{SQ}\right\}.$$

References

Baillie, R. T., Bollerslev, T. & Mikkelsen, H. O. (1996), 'Fractionally integrated generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics* **74**, 3–30.

Barndorff-Nielsen, O. E. & Shephard, N. (2002), 'Estimating quadratic variation using realised volatility', *Journal of Applied Econometrics* **17**, 457–477.

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

Black, F. (1976), 'Studies in stock price volatility changes'. Proceedings of the 1976 business meeting of the business and economics section, American Statistical Association, 177-181.

Bollerslev, T. (1986), 'Generalized autoregressive heteroskedasticity', *Journal of Econometrics* **31**, 307–327.

Clark, T. E. & McCracken, M. W. (2001), 'Tests of equal forecast accuracy and encompassing for nested models', *Journal of Econometrics* **105**, 85–110.

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

Ding, Z., Granger, C. W. J. & Engle, R. F. (1993), 'A long memory property of stock market returns and a new model', *Journal of Empirical Finance* **1**, 83–106.

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

Elton, E. J., Gruber, M. J., Comer, G. & Li, K. (2002), 'Spiders : Where are the bugs?', *Journal of Business* **75**(3), 453–472.

Engle, R. F. (1982), 'Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation', *Econometrica* **45**, 987–1007.

Engle, R. F. & Ng, V. (1993), 'Measuring and testing the impact of news on volatility', *Journal of Finance* **48**, 1747–1778.

Geweke, J. (1986), 'Modelling persistence in conditional variances: A comment', *Econometric Reviews* **5**, 57–61.

Glosten, L. R., Jagannathan, R. & Runkle, D. E. (1993), 'On the relation between the expected value and the volatility of the nominal excess return on stocks', *Journal of Finance* **48**, 1779–1801.

Hansen, P. R. (2001), 'A test for superior predictive ability'. Brown University, Economics Working Paper. 2001-06
`http://www.econ.brown.edu/fac/Peter_Hansen`.

Hansen, P. R. (2003*a*), 'Asymptotic tests of composite hypotheses'. Brown University, Department of Economics Working Paper, 2003-09.
`http://www.econ.brown.edu/fac/Peter_Hansen`.

Hansen, P. R. (2003*b*), 'Regression analysis with many specifications: A bootstrap method to robust inference'. Mimeo
`http://www.econ.brown.edu/fac/Peter_Hansen`.

Hansen, P. R. & Lunde, A. (2001), 'A forecast comparison of volatility models: Does anything beat a GARCH(1,1)?'. Brown University Working Paper, 2001-04
`http://www.econ.brown.edu/fac/Peter_Hansen`.

Hansen, P. R. & Lunde, A. (2003), 'Consistent preordering with an estimated criterion function, with an application to the evaluation and comparison of volatility models'. Brown University Working Paper, 2003-01
`http://www.econ.brown.edu/fac/Peter_Hansen`.

Hansen, P. R., Lunde, A. & Nason, J. M. (2003), 'Model confidence sets for forecasting models'. Mimeo.

Härdle, W., Horowitz, J. & Kreiss, J.-P. (2002), 'Bootstrap methods for time series', *Mimeo* .

Hasbrouck, J. (2002), Intra day price formation in US equity markets. Manuscript, NYU.

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

Higgins, M. L. & Bera, A. K. (1992), 'A class of nonlinear ARCH models', *International Economic Review* **33**, 137–158.

Johansen, S. (1988), 'Statistical analysis of cointegration vectors', *Journal of Economic Dynamics and Control* **12**, 231–254.

Johansen, S. (1996), *Likelihood Based Inference in Cointegrated Vector Autoregressive Models*, 2nd edn, Oxford University Press, Oxford.

Koopman, S., Shephard, N. & J.A.Doornik (1999), 'Statistical algorithms for models in state space form using ssfpack 2.2', *Econometrics Journal* **2**, 113–166.

Loudon, G. F., Watt, W. H. & Yadav, P. K. (2000), 'An empirical analysis of alternative parametric ARCH models', *Journal of Applied Econometrics* **15**(1), 117–136.

Malliavin, P. & Mancino, M. E. (2002), 'Fourier series method for measurement of multivariate volatilities', *Finance and Stochastics* **6**(1), 49–61.

Nelson, D. B. (1991), 'Conditional heteroskedasticity in asset retruns: A new approach', *Econometrica* **59**, 347–370.

Pantula, S. G. (1986), 'Modelling persistence in conditional variances: A comment', *Econometric Reviews* **5**, 71–74.

Pantula, S. G. (1989), 'Testing for unit roots in time series data', *Econometric Theory* **5**, 256–71.

Poon, S.-H. & Granger, C. (2003), 'Forecasting volatility in financial markets: A review', *Journal of Economic Literature* **41**, 478–539.

Schwert, G. W. (1989), 'Why does stock volatility change over time?', *Journal of Finance* **44**(5), 1115–1153.

Sentana, E. (1995), 'Quadratic ARCH models', *Review of Economic Studies* **62**(4), 639–661.

Taylor, S. J. (1986), *Modelling Financial Time Series*, John Wiley & Sons, NY.

West, K. D. (1996), 'Asymptotic inference about predictive ability', *Econometrica* **64**, 1067–1084.

West, K. D. & McCracken, M. W. (1998), 'Regression based tests of predictive ability', *International Economic Review* **39**, 817–840.

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 I
Alternative GARCH-type models: The conditional variance

SV:	$\ln(\sigma_t^2) = \omega + \beta \ln(\sigma_{t-1}^2) + \eta_t, \quad \eta_t \sim Nid(0, \sigma_\eta^2)$
ARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2$
GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
IGARCH	$\sigma_t^2 = \omega + \varepsilon_{t-1}^2 + \sum_{i=2}^{q} \alpha_i (\varepsilon_{t-i}^2 - \varepsilon_{t-1}^2) + \sum_{j=1}^{p} \beta_j (\sigma_{t-j}^2 - \varepsilon_{t-1}^2)$
Taylor/Schwert:	$\sigma_t = \omega + \sum_{i=1}^{q} \alpha_i \lvert\varepsilon_{t-i}\rvert + \sum_{j=1}^{p} \beta_j \sigma_{t-j}$
A-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \left[\alpha_i \varepsilon_{t-i}^2 + \gamma_i \varepsilon_{t-i}\right] + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
NA-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \left(\varepsilon_{t-i} + \gamma_i \sigma_{t-i}\right)^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
V-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \left(e_{t-i} + \gamma_i\right)^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
Thr.-GARCH:	$\sigma_t = \omega + \sum_{i=1}^{q} \alpha_i \left[(1-\gamma_i)\varepsilon_{t-i}^{+} - (1+\gamma_i)\varepsilon_{t-i}^{-}\right] + \sum_{j=1}^{p} \beta_j \sigma_{t-j}$
GJR-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \left[\alpha_i + \gamma_i I_{\{\varepsilon_{t-i}>0\}}\right] \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
log-GARCH:	$\ln(\sigma_t) = \omega + \sum_{i=1}^{q} \alpha_i \lvert e_{t-i}\rvert + \sum_{j=1}^{p} \beta_j \ln(\sigma_{t-j})$
EGARCH:	$\ln(\sigma_t^2) = \omega + \sum_{i=1}^{q} \left[\alpha_i e_{t-i} + \gamma_i \left(\lvert e_{t-i}\rvert - E\lvert e_{t-i}\rvert\right)\right] + \sum_{j=1}^{p} \beta_j \ln(\sigma_{t-j}^2),$
A-PARCH:	$\sigma^\delta = \omega + \sum_{i=1}^{q} \alpha_i \left[\lvert\varepsilon_{t-i}\rvert - \gamma_i \varepsilon_{t-i}\right]^\delta + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^\delta$
GQ-ARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i} + \sum_{i=1}^{q} \alpha_{ii} \varepsilon_{t-i}^2 + \sum_{i<j}^{q} \alpha_{ij} \varepsilon_{t-i} \varepsilon_{t-j} + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
FI-GARCH:	$\sigma_t^2 = \omega[1-\beta(L)]^{-1} + \left\{1 - [1-\beta(L)]^{-1}\phi(L)(1-L)^d\right\} \varepsilon_t^2$
FI-EGARCH:*	$\ln(\sigma_t^2) = \omega + \phi(L)^{-1}(1-L)^{-d}[1-\alpha(L)]g(e_{t-1})$

Notes: *Here $g(z_t) = \gamma_1 z_t + \gamma_2 \left(\lvert z_t\rvert - E\lvert z_t\rvert\right)$.

TABLE II
MCS and p-values

No.	Model	MSE	$\widehat{var}(\bar{d}_{i.})$	p_R	p_{SQ}	MAD	$\widehat{var}(\bar{d}_{i.})$	p_R	p_{SQ}
1	STOCH.VO	1.617	3.887	0.444⋆⋆	0.176⋆	0.698	0.150	0.054	0.000
2	GARCH11	1.703	1.712	0.020	0.029	0.765	0.089	0.002	0.000
3	GARCH12	1.825	2.964	0.020	0.004	0.789	0.106	0.002	0.000
4	GARCH21	1.737	1.935	0.020	0.008	0.772	0.090	0.002	0.000
5	GARCH22	1.858	3.435	0.020	0.003	0.794	0.137	0.002	0.000
6	IGARCH11	1.731	1.708	0.020	0.010	0.779	0.098	0.002	0.000
7	IGARCH12	1.868	2.856	0.020	0.002	0.808	0.120	0.002	0.000
8	IGARCH21	1.771	1.904	0.020	0.006	0.787	0.101	0.002	0.000
9	IGARCH22	1.916	3.405	0.020	0.001	0.820	0.153	0.002	0.000
10	TSGARCH11	1.679	2.044	0.078	0.073	0.769	0.108	0.004	0.000
11	TSGARCH12	1.739	2.557	0.020	0.023	0.776	0.107	0.002	0.000
12	TSGARCH21	1.695	2.142	0.078	0.038	0.770	0.106	0.004	0.000
13	TSGARCH22	1.759	2.607	0.020	0.008	0.781	0.115	0.002	0.000
14	AGARCH11	1.584	0.809	0.444⋆⋆	0.198⋆	0.727	0.026	0.054	0.000
15	AGARCH12	1.547	1.430	0.934⋆⋆	0.916⋆⋆	0.704	0.059	0.054	0.006
16	AGARCH21	1.575	0.774	0.444⋆⋆	0.274⋆⋆	0.725	0.026	0.054	0.000
17	AGARCH22	1.704	2.110	0.078	0.032	0.744	0.071	0.054	0.000
18	NAGARCH11	1.878	3.376	0.020	0.006	0.812	0.116	0.002	0.000
19	NAGARCH12	1.852	2.966	0.020	0.008	0.804	0.105	0.002	0.000
20	NAGARCH21	1.873	4.037	0.020	0.060	0.817	0.123	0.002	0.000
21	NAGARCH22	2.012	13.81	0.444⋆⋆	0.104⋆	0.880	0.330	0.002	0.000
22	VGARCH11	1.587	7.203	0.594⋆⋆	0.247⋆	0.659	0.231	0.054	0.059
23	VGARCH12	1.578	6.673	0.630⋆⋆	0.312⋆⋆	0.653	0.240	1.000⋆⋆	1.000⋆⋆
24	VGARCH21	1.583	6.971	0.630⋆⋆	0.312⋆⋆	0.659	0.230	0.054	0.059
25	VGARCH22	1.527	4.960	1.000⋆⋆	1.000⋆⋆	0.665	0.173	0.054	0.059
26	ThrGARCH11	1.823	2.293	0.020	0.007	0.815	0.083	0.002	0.000
27	ThrGARCH12	1.810	3.411	0.444⋆⋆	0.036	0.831	0.126	0.002	0.000
28	ThrGARCH21	1.816	2.566	0.020	0.010	0.817	0.084	0.002	0.000
29	ThrGARCH22	1.707	5.481	0.444⋆⋆	0.176⋆	0.825	0.204	0.002	0.000
30	GJRGARCH11	1.803	2.092	0.020	0.008	0.777	0.114	0.054	0.000
31	GJRGARCH21	1.803	2.082	0.020	0.007	0.776	0.114	0.054	0.000
32	logGARCH11	1.622	2.035	0.444⋆⋆	0.176⋆	0.740	0.088	0.054	0.000
33	logGARCH12	1.652	2.417	0.444⋆⋆	0.096	0.741	0.082	0.054	0.000
34	logGARCH21	1.631	2.129	0.444⋆⋆	0.176⋆	0.740	0.085	0.054	0.000
35	logGARCH22	1.686	2.602	0.444⋆⋆	0.052	0.748	0.089	0.054	0.000
36	EGARCH11	1.651	0.795	0.444⋆⋆	0.176⋆	0.750	0.048	0.054	0.000
37	EGARCH12	1.685	0.942	0.078	0.052	0.769	0.089	0.004	0.000
38	EGARCH21	1.644	0.846	0.444⋆⋆	0.176⋆	0.749	0.049	0.054	0.000
39	EGARCH22	1.551	2.708	0.934⋆⋆	0.916⋆⋆	0.753	0.149	0.054	0.000
40	APARCH11	1.902	2.564	0.020	0.002	0.854	0.117	0.002	0.000
41	APARCH12	1.806	1.954	0.020	0.006	0.821	0.098	0.002	0.000
42	APARCH21	1.872	2.518	0.020	0.004	0.847	0.109	0.002	0.000
43	APARCH22	1.803	1.825	0.020	0.006	0.818	0.097	0.002	0.000
44	GQARCH11	1.584	0.809	0.444⋆⋆	0.176⋆	0.727	0.026	0.054	0.000
45	GQARCH12	1.647	1.520	0.444⋆⋆	0.176⋆	0.765	0.106	0.054	0.000
46	GQARCH21	1.575	0.774	0.630⋆⋆	0.312⋆⋆	0.725	0.026	0.054	0.000
47	GQARCH22	1.586	1.528	0.630⋆⋆	0.312⋆⋆	0.740	0.128	0.054	0.000
48	FIGARCH00	1.717	5.128	0.444⋆⋆	0.088	0.735	0.260	0.054	0.000
49	FIGARCH01	1.905	7.772	0.020	0.007	0.781	0.210	0.054	0.000
50	FIGARCH10	1.840	4.752	0.020	0.006	0.777	0.124	0.002	0.000
51	FIGARCH11	1.853	5.290	0.020	0.006	0.778	0.138	0.004	0.000
52	FIEGARCH00	2.606	28.99	0.020	0.012	1.112	0.577	0.002	0.000
53	FIEGARCH01	2.516	17.99	0.020	0.002	1.085	0.520	0.002	0.000
54	FIEGARCH10	2.535	23.52	0.020	0.004	1.092	0.537	0.002	0.000
55	FIEGARCH11	2.538	19.31	0.020	0.002	1.091	0.532	0.002	0.000

Notes: This table present model confidence sets for our selection of volatility models. P-values marked with two stars are in $\widehat{\mathcal{M}}^*_{0.25}$, and one star marks models $\widehat{\mathcal{M}}^*_{0.1}$. Note that $\widehat{\mathcal{M}}^*_{0.25} \subset \widehat{\mathcal{M}}^*_{0.1}$

TABLE III
Decomposition of loss function variance.

Model	MSE	$\frac{\text{MSE}}{\sqrt{\widehat{V}_{\bar{d}_{i,.}}}}$	$\sqrt{\widehat{V}_{\bar{d}_{i,.}}}$	$\frac{\widehat{V}_{\bar{L}_{i,.}}}{\widehat{V}_{\bar{L}_{.,.}}}$	$\hat{\rho}_{\bar{L}_{i,.},\bar{L}_{.,.}}$	MAD	$\frac{\text{MAD}}{\sqrt{\widehat{V}_{\bar{d}_{i,.}}}}$	$\sqrt{\widehat{V}_{\bar{d}_{i,.}}}$	$\frac{\widehat{V}_{\bar{L}_{i,.}}}{\widehat{V}_{\bar{L}_{.,.}}}$	$\hat{\rho}_{\bar{L}_{i,.},\bar{L}_{.,.}}$
STOCH.VO	1.617	0.820	1.971	1.192	0.994	0.698	1.803	0.387	1.127	0.931
GARCH11	1.703	1.301	1.309	1.068	0.995	0.765	2.563	0.299	1.113	0.959
GARCH12	1.825	1.060	1.722	1.169	0.995	0.789	2.419	0.326	1.200	0.955
GARCH21	1.737	1.249	1.391	1.097	0.996	0.772	2.566	0.301	1.138	0.959
GARCH22	1.858	1.002	1.853	1.165	0.994	0.794	2.142	0.371	1.224	0.943
IGARCH11	1.731	1.324	1.307	1.051	0.995	0.779	2.482	0.314	1.121	0.955
IGARCH12	1.868	1.106	1.690	1.144	0.995	0.808	2.335	0.346	1.212	0.950
IGARCH21	1.771	1.283	1.380	1.078	0.995	0.787	2.474	0.318	1.147	0.955
IGARCH22	1.916	1.038	1.845	1.133	0.993	0.820	2.094	0.392	1.239	0.937
TSGARCH11	1.679	1.175	1.430	1.045	0.994	0.769	2.339	0.329	1.085	0.949
TSGARCH12	1.739	1.087	1.599	1.106	0.994	0.776	2.370	0.328	1.133	0.951
TSGARCH21	1.695	1.158	1.463	1.063	0.994	0.770	2.362	0.326	1.099	0.950
TSGARCH22	1.759	1.089	1.615	1.104	0.994	0.781	2.300	0.339	1.146	0.948
AGARCH11	1.584	1.761	0.899	1.053	0.998	0.727	4.521	0.161	1.052	0.987
AGARCH12	1.547	1.294	1.196	1.082	0.996	0.704	2.909	0.242	1.048	0.971
AGARCH21	1.575	1.790	0.880	1.043	0.998	0.725	4.480	0.162	1.046	0.987
AGARCH22	1.704	1.173	1.453	1.149	0.997	0.744	2.790	0.267	1.146	0.969
NAGARCH11	1.878	1.022	1.838	0.943	0.991	0.812	2.387	0.340	1.214	0.952
NAGARCH12	1.852	1.075	1.722	0.950	0.992	0.804	2.485	0.324	1.201	0.956
NAGARCH21	1.873	0.932	2.009	0.898	0.989	0.817	2.328	0.351	1.202	0.948
NAGARCH22	2.012	0.541	3.716	0.813	0.964	0.880	1.531	0.575	1.234	0.857
VGARCH11	1.587	0.591	2.684	1.308	0.992	0.659	1.372	0.481	1.149	0.895
VGARCH12	1.578	0.611	2.583	1.283	0.991	0.653	1.331	0.490	1.148	0.890
VGARCH21	1.583	0.600	2.640	1.299	0.992	0.659	1.376	0.479	1.145	0.895
VGARCH22	1.527	0.686	2.227	1.228	0.993	0.665	1.598	0.416	1.082	0.917
ThrGARCH11	1.823	1.204	1.514	0.911	0.994	0.815	2.832	0.288	1.156	0.964
ThrGARCH12	1.810	0.980	1.847	0.830	0.994	0.831	2.339	0.356	1.116	0.941
ThrGARCH21	1.816	1.133	1.602	0.887	0.994	0.817	2.813	0.290	1.145	0.962
ThrGARCH22	1.707	0.729	2.341	0.762	0.992	0.825	1.827	0.452	1.023	0.899
GJRGARCH11	1.803	1.247	1.446	0.958	0.994	0.777	2.301	0.337	1.196	0.952
GJRGARCH21	1.803	1.249	1.443	0.960	0.994	0.776	2.299	0.338	1.196	0.952
logGARCH11	1.622	1.137	1.427	1.083	0.995	0.740	2.499	0.296	1.071	0.958
logGARCH12	1.652	1.063	1.555	1.126	0.995	0.741	2.587	0.287	1.099	0.962
logGARCH21	1.631	1.118	1.459	1.097	0.995	0.740	2.539	0.292	1.079	0.959
logGARCH22	1.686	1.045	1.613	1.137	0.995	0.748	2.504	0.299	1.123	0.959
EGARCH11	1.651	1.851	0.892	0.977	0.997	0.750	3.407	0.220	1.086	0.977
EGARCH12	1.685	1.735	0.971	0.962	0.997	0.769	2.574	0.299	1.090	0.958
EGARCH21	1.644	1.787	0.920	0.965	0.997	0.749	3.400	0.220	1.079	0.977
EGARCH22	1.551	0.943	1.646	0.842	0.995	0.753	1.951	0.386	0.982	0.925
APARCH11	1.902	1.188	1.601	0.917	0.993	0.854	2.494	0.342	1.169	0.948
APARCH12	1.806	1.292	1.398	0.909	0.995	0.821	2.626	0.313	1.129	0.955
APARCH21	1.872	1.180	1.587	0.898	0.994	0.847	2.565	0.330	1.150	0.951
APARCH22	1.803	1.334	1.351	0.919	0.995	0.818	2.629	0.311	1.130	0.956
GQARCH11	1.584	1.761	0.899	1.053	0.998	0.727	4.521	0.161	1.052	0.987
GQARCH12	1.647	1.336	1.233	0.948	0.996	0.765	2.348	0.326	1.059	0.949
GQARCH21	1.575	1.790	0.880	1.043	0.998	0.725	4.480	0.162	1.046	0.987
GQARCH22	1.586	1.283	1.236	0.957	0.995	0.740	2.066	0.358	1.035	0.937
FIGARCH00	1.717	0.758	2.265	1.005	0.986	0.735	1.441	0.510	1.173	0.883
FIGARCH01	1.905	0.684	2.788	1.303	0.990	0.781	1.706	0.458	1.292	0.916
FIGARCH10	1.840	0.844	2.180	1.265	0.995	0.777	2.204	0.353	1.232	0.949
FIGARCH11	1.853	0.806	2.300	1.279	0.994	0.778	2.095	0.371	1.244	0.944
FIEGARCH00	2.606	0.484	5.385	0.671	0.925	1.112	1.464	0.760	1.365	0.766
FIEGARCH01	2.516	0.593	4.242	0.751	0.954	1.085	1.505	0.721	1.336	0.786
FIEGARCH10	2.535	0.523	4.850	0.691	0.941	1.092	1.489	0.733	1.339	0.779
FIEGARCH11	2.538	0.578	4.394	0.748	0.950	1.091	1.496	0.729	1.344	0.782

Notes: This table present some summary statistics of the loss function variance decomposition.

TABLE IV
Simulations with m^* superior (EPA) models and $m - m^*$ inferior models, $\rho = m^*/m = 0.5$.

Panel A: $\alpha = 0.1$

λ/m		10			40		
		T_{Bonf}	T_R	T_{SQ}	T_{Bonf}	T_R	T_{SQ}
1	$Q_{\mathcal{M}^*\subset\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.944	0.948	1.000	0.936	0.909
	$Q_{\mathcal{M}^*=\widehat{\mathcal{M}}^*_\alpha}$	0.000	0.000	0.001	0.000	0.000	0.000
	$Q_{\mathcal{M}^*\mid\widehat{\mathcal{M}}^*_\alpha}$	0.500	0.517	0.521	0.500	0.507	0.512
	$Q_{\complement\mathcal{M}^*\mid\complement\widehat{\mathcal{M}}^*_\alpha}$	-	0.855	0.876	-	0.897	0.912
5	$Q_{\mathcal{M}^*\subset\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.898	0.892	1.000	0.892	0.875
	$Q_{\mathcal{M}^*=\widehat{\mathcal{M}}^*_\alpha}$	0.000	0.782	0.810	0.000	0.367	0.384
	$Q_{\mathcal{M}^*\mid\widehat{\mathcal{M}}^*_\alpha}$	0.500	0.977	0.986	0.500	0.938	0.971
	$Q_{\complement\mathcal{M}^*\mid\complement\widehat{\mathcal{M}}^*_\alpha}$	-	0.978	0.976	-	0.990	0.992
20	$Q_{\mathcal{M}^*\subset\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.893	0.893	1.000	0.898	0.896
	$Q_{\mathcal{M}^*=\widehat{\mathcal{M}}^*_\alpha}$	0.000	0.893	0.893	0.000	0.898	0.896
	$Q_{\mathcal{M}^*\mid\widehat{\mathcal{M}}^*_\alpha}$	0.500	1.000	1.000	0.500	1.000	1.000
	$Q_{\complement\mathcal{M}^*\mid\complement\widehat{\mathcal{M}}^*_\alpha}$	-	0.977	0.976	-	0.990	0.992
40	$Q_{\mathcal{M}^*\subset\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.893	0.898	1.000	0.888	0.906
	$Q_{\mathcal{M}^*=\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.893	0.898	0.014	0.888	0.906
	$Q_{\mathcal{M}^*\mid\widehat{\mathcal{M}}^*_\alpha}$	1.000	1.000	1.000	0.510	1.000	1.000
	$Q_{\complement\mathcal{M}^*\mid\complement\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.978	0.978	1.000	0.990	0.993

Panel B: $\alpha = 0.05$

λ/m		10			40		
		T_{Bonf}	T_R	T_{SQ}	T_{Bonf}	T_R	T_{SQ}
1	$Q_{\mathcal{M}^*\subset\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.972	0.975	1.000	0.973	0.952
	$Q_{\mathcal{M}^*=\widehat{\mathcal{M}}^*_\alpha}$	0.000	0.000	0.000	0.000	0.000	0.000
	$Q_{\mathcal{M}^*\mid\widehat{\mathcal{M}}^*_\alpha}$	0.500	0.509	0.511	0.500	0.503	0.507
	$Q_{\complement\mathcal{M}^*\mid\complement\widehat{\mathcal{M}}^*_\alpha}$	-	0.867	0.898	-	0.921	0.926
5	$Q_{\mathcal{M}^*\subset\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.952	0.946	1.000	0.943	0.929
	$Q_{\mathcal{M}^*=\widehat{\mathcal{M}}^*_\alpha}$	0.000	0.750	0.787	0.000	0.271	0.277
	$Q_{\mathcal{M}^*\mid\widehat{\mathcal{M}}^*_\alpha}$	0.500	0.958	0.972	0.500	0.902	0.959
	$Q_{\complement\mathcal{M}^*\mid\complement\widehat{\mathcal{M}}^*_\alpha}$	-	0.990	0.989	-	0.995	0.996
20	$Q_{\mathcal{M}^*\subset\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.942	0.946	1.000	0.945	0.948
	$Q_{\mathcal{M}^*=\widehat{\mathcal{M}}^*_\alpha}$	0.000	0.942	0.946	0.000	0.945	0.948
	$Q_{\mathcal{M}^*\mid\widehat{\mathcal{M}}^*_\alpha}$	0.500	1.000	1.000	0.500	1.000	1.000
	$Q_{\complement\mathcal{M}^*\mid\complement\widehat{\mathcal{M}}^*_\alpha}$	-	0.988	0.989	-	0.995	0.996
40	$Q_{\mathcal{M}^*\subset\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.944	0.942	1.000	0.944	0.952
	$Q_{\mathcal{M}^*=\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.944	0.942	0.000	0.944	0.952
	$Q_{\mathcal{M}^*\mid\widehat{\mathcal{M}}^*_\alpha}$	1.000	1.000	1.000	0.500	1.000	1.000
	$Q_{\complement\mathcal{M}^*\mid\complement\widehat{\mathcal{M}}^*_\alpha}$	1.000	0.989	0.988	1.000	0.996	0.997

Notes: This table present the Bonferoni simulation set for $\rho = 0.5$.

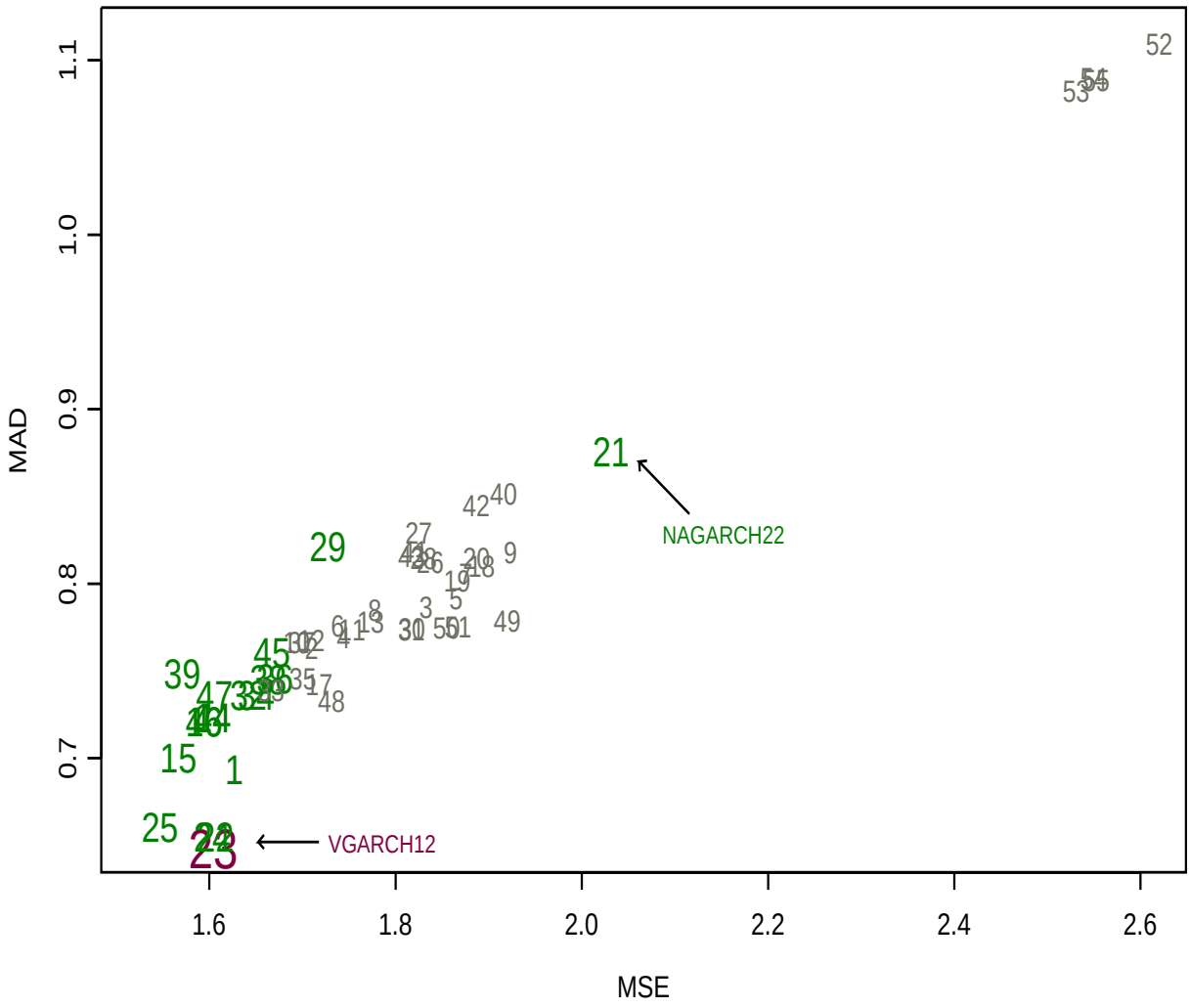

Figure 1: This figure present the volatility $\widehat{\mathcal{M}}^{*}_{10\%}$ based on sequential testing using the T_{SQ} statistics. The models are plotted according to their MAD/MSE combination. The model number in larger boldface are in $\widehat{\mathcal{M}}^{*}_{10\%}$ both according to MSE and MAD loss function. The slightly smaller regular faced number are in $\widehat{\mathcal{M}}^{*}_{10\%}$ according to the MSE loss. The small numbers are excluded from the $\widehat{\mathcal{M}}^{*}_{10\%}$.

Estimated Classification		Population Classification: superior: $\mathcal{M}^*$	Population Classification: inferior: $\complement\mathcal{M}^*$
	superior: $\widehat{\mathcal{M}}_\alpha^*$	A_{11} # Correctly classified as superior model	A_{12} # Inferior models wrongly classified as superior models
	inferior: $\complement\widehat{\mathcal{M}}_\alpha^*$	A_{21} # Superior models wrongly classified as inferior models	A_{22} # Correctly classified as inferior models

Figure 2: Classification into superior and inferior model