Model Selection and Testing of Conditional and Stochastic Volatility Models*

Massimiliano Caporin
Department of Economics and Management "Marco Fanno"
University of Padova
Italy

Michael McAleer
Econometric Institute
Erasmus School of Economics
Erasmus University Rotterdam
and
Timbergen Institute
The Netherlands
and
Institute of Economic Research
Kyoto University
Japan

EI 2010-57

September 2010

* The second author wishes to acknowledge the financial support of the Australian Research Council, National Science Council, Taiwan, and the Japan Society for the Promotion of Science. This paper is to appear in *Handbook on Financial Engineering and Econometrics: Volatility Models and Their Applications*, L. Bauwens, C. Hafner and S. Laurent (eds.), Wiley, New York.

Abstract

This paper focuses on the selection and comparison of alternative non-nested volatility models. We review the traditional in-sample methods commonly applied in the volatility framework, namely diagnostic checking procedures, information criteria, and conditions for the existence of moments and asymptotic theory, as well as the out-of-sample model selection approaches, such as mean squared error and Model Confidence Set approaches. The paper develops some innovative loss functions which are based on Value-at-Risk forecasts. Finally, we present an empirical application based on simple univariate volatility models, namely GARCH, GJR, EGARCH, and Stochastic Volatility that are widely used to capture asymmetry and leverage.

Keywords: Volatility model selection, volatility model comparison, non-nested models, model confidence set, Value-at-Risk forecasts, asymmetry, leverage.

JEL: C11, C22, C52, C58.

1. Introduction

Model selection and model comparison, especially of the conditional mean or first moment of a given random variable, has been widely considered in the sciences and social sciences for an extended period. The relevance and importance of such a topic comes from the recognized fact that the true data generating processes are generally unknown.

As a result, several approaches have been proposed to verify if a given model is able to replicate or capture the empirical features observed on sample data (the realization of the data generating process), and to check if there is a preference across alternative models that might be considered given the sample data and the purposes of the analysis.

In this paper we focus on model comparison and selection in a specific framework, namely univariate volatility models for financial time series. From the seminal work of Engle (1982) and Bollerslev (1986), GARCH models have become a very popular tool in empirical finance. They have been generalized in several ways (see, for example, Bollerslev et al. (1992, 1994) and McAleer (2005)). A companion family of models is that of stochastic volatility (SV), introduced by Taylor (1982, 1986), and extended in several directions (see, for example, Ghysels et al. (1996) and Asai et al. (2006)).

Traditional methods for model selection and comparison could easily be extended and applied within specific families of models (for instance, within GARCH or within SV specifications). However, some model classes, or some specific models within a given model class, may be non-nested, thereby requiring appropriate approaches or novel techniques for the model selection step.

In the following, we will consider separately the comparison of alternative specifications in-sample, thereby resorting to nested and non-nested model comparison, diagnostic checking, and out-of-sample model comparison based on the forecasts of given models.

The discussion herein is based on univariate models that are capable of capturing financial time series asymmetry and/or leverage, but the results presented can be generalized to other model classes at the univariate level. The methods can also be generalized to the multivariate level, following Patton and Sheppard (2009) and Caporin and McAleer (2009, 2010).

The remainder of the paper proceeds as follow. In subsection 1.1, we introduce the models to be used. Section 2 discusses the model selection and testing methods, distinguishing between in-sample and out-of-sample approaches. Section 3 includes an empirical example on a set of stock market indices. Finally, Section 4 gives some concluding comments.

1.1 Model specifications

In this paper we illustrate some approaches to model selection and comparison making use of simple and well-known univariate volatility models. We consider the traditional GARCH(1,1), its extension to capture asymmetry, the GJR(1,1) model of Glosten et al. (2002), Exponential GARCH(1,1) (EGARCH), and the Autoregressive SV(1) (also known as SV) specifications. We choose these three models for two simple reasons, namely they are non-nested and can capture asymmetry and leverage (with the obvious exclusion of GARCH(1,1), which is a benchmark model).

In order to simplify model evaluation and comparison, we assume in the following that the analyzed return series, r_t , has been filtered from its mean, so that we can focus on a zero-mean series, ε_t , that display conditional heteroskedasticity, $\varepsilon_t = \sigma_t z_t$. Furthermore, the unit variance innovation, z_t , is a standardized residual.

If the conditional variances, σ_t^2 , follow a GARCH(1,1) model, the following equation represents their law of motion:

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{1}$$

where $\omega \geq 0$, $\alpha \geq 0$, $\beta \geq 0$, and $\alpha + \beta \leq 1$ are sufficient conditions to guarantee positive conditional variances for all observations.

If we introduce asymmetry to GARCH(1,1), we obtain the asymmetric or threshold model of Glosten et al. (1993), GJR(1,1):

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 I\left(\varepsilon_{t-1} < 0\right) + \beta\sigma_{t-1}^2 \tag{2}$$

Where the parameter γ captures asymmetry, and $I\left(\varepsilon_{t-1} < 0\right)$ is an indicator function, which takes the value 1 when $\varepsilon_{t-1} < 0$, and 0 otherwise.

A clarification is needed here in order to avoid a common misconception between asymmetry and leverage: (i) asymmetry is a feature that is intended to capture the empirical regularity that positive and negative shocks of equal magnitude have different impacts on volatility; (ii)

leverage is intended to capture the possibility that negative shocks increase volatility while positive shocks decrease volatility.

As a matter of model design, few conditional volatility models allow for leverage effects. For example, GARCH is symmetric and hence has no leverage. Despite comments to the contrary in various econometric software packages (for instance, EViews and Matlab), the GJR model (also known as Threshold GARCH, or TGARCH) may be asymmetric, but it is unlikely to have leverage, as the ARCH effect must be negative, which is contrary to virtually every empirical finding in the financial econometrics literature.

The third model we consider is the EGARCH(1,1), where the conditional variance equation is defined in term of log-variances:

$$\ln\left(\sigma_t^2\right) = \omega + \alpha\left|z_{t-1}\right| + \gamma z_{t-1} + \beta \ln\left(\sigma_{t-1}^2\right). \tag{3}$$

Note that the coefficients need not to be positive, while $|\beta| < 1$ to avoid explosive variance patterns. In equation (3), the parameters α and γ influence the presence of asymmetry and leverage: if $z_{t-1} \geq 0$, the shock's impact on conditional variances is $(\alpha + \gamma) z_{t-1}$, while if $z_{t-1} < 0$, the impact is $(\gamma - \alpha) z_{t-1}$. As a result, if $\gamma = 0$ the model is symmetric, and hence cannot have leverage. We have asymmetry (with a larger impact on volatility of negative shocks as compared with positive shocks of similar magnitude) if $\gamma < 0$, and leverage (whereby negative shocks increase volatility and positive shocks decrease volatility) if $\gamma < 0$ and $\gamma < \alpha < -\gamma$. The EGARCH model can have leverage, but two restrictions on the parameters α and γ must be satisfied (some econometric software manuals state incorrectly that leverage arises through a constraint on a single parameter, namely γ).

Finally, we consider the stochastic volatility model, which assumes that the innovation term follows:

$$\varepsilon_t = \exp\left(\frac{1}{2} h_t\right) z_t \tag{4}$$

where z_t is a unit variance innovation and the conditional variance $\exp(h_t)$ is driven by the following dynamic equation for h_t:

$$h_{t+1} = \phi_0 + \phi_1 h_t + \eta_t \qquad (5)$$

where the parameters are not required to be positive, $|\phi_1| < 1$ to avoid explosive patterns, and the innovation term, η_t , has variance σ_η^2 .

As shown Yu (2005), the SV model displays a leverage effect if the two innovation terms, η_t and z_t , are negatively correlated, while asymmetry may be included following, for instance, the approaches of Danielsson (1994), So et al. (2002), and Asai and McAleer (2005).

2. Model Selection and Testing

Selection of the best or the most appropriate model may be based on in-sample or out-of-sample criteria, or both. In the following, we will address these two approaches separately. Such a choice derives from purely illustrative purposes, and should not be interpreted as a preference for one of the two methods. Indeed, identification of an optimal model would seem to require an optimal balance between these two approaches.

In empirical applications we search for models that capture the features of the analyzed data, and that provide accurate out-of-sample forecasts. Both elements may not be present over all models and thus, in empirical studies, a trade-off will likely exist. This possible inconsistency may be resolved in part by evaluating the purpose of an empirical exercise. Structural analysis may have greater emphasis on in-sample fit, while forecasting exercises will necessarily concentrate on out-of-sample outcomes. Nevertheless, both aspects need to be considered, as does the role of research expertise.

2.1 In-sample comparisons

This paper examines conditional volatility (GARCH) models and stochastic volatility (SV) processes. We focus on alternative model specifications that belong to the same family (either GARCH or SV). If the models we compare have known mathematical and asymptotic properties (such as strict stationarity of the underlying random process, and consistency and asymptotic normality of the estimators), we may compare them by checking if the conditions ensuring the existence of moments or asymptotic properties are satisfied. In principle, models

where these conditions are not satisfied, or do not even exist, should be discarded. In practice, this is typically not the case.

For instance, log-moment conditions ensuring strict stationarity and ergodicity of GARCH models are reported in Nelson (1990) and Bougerol and Picard (1992), among others. These conditions are also sufficient for consistency and asymptotic normality of quasi-maximum likelihood estimators (QMLE) (for example, see Elie and Jeantheau (1995), and Boussama (2000)). Stronger but simpler moment conditions for ergodicity, stationarity, consistency and asymptotic normality of the QMLE, have been provided in Ling and McAleer (2002a, b) and McAleer et al. (2007). In practice, log-moment conditions are generally difficult to verify, especially for multivariate processes, while moment conditions may be considered as a useful diagnostic check. Notably, well written software should implement these conditions (which are generally represented as non-linear parametric restrictions) within the estimation step, thereby implicitly checking them.

As an example, consider the GJR model of equation (2). In this case, the stationarity and ergodicity condition, under the assumption that shocks follow a symmetric density, is given as $\alpha + 0.5\gamma + \beta < 1$, while the condition for the existence of the fourth-order moment is $\beta^2 + 2\alpha\beta + 3\alpha^2 + \beta\gamma + 3\alpha\gamma + 0.5\gamma^2 < 1$. The log-moment condition is given as $E\left[\ln\left(\alpha z_{t-1}^2 + \gamma z_{t-1}^2 I\left(z_{t-1} < 0\right) + \beta\right)\right] < 0$, but it could be difficult to verify as it requires the evaluation of the expectation of a function of an unknown random variance and of unknown coefficients.

From a different viewpoint, we may compare models with respect to the features they are supposed to be capturing. For example, we may prefer volatility models with asymmetry to specifications characterized by a symmetric news impact curve.

Model preference based on model flexibility should obviously be matched with the statistical significance of estimated parameters associated with a particular feature. For instance, referring to the GJR model, it can capture asymmetry though not leverage, and hence is more flexible than the symmetric GARCH(1,1) specification. However, GJR should be preferred empirically to GARCH if the estimated asymmetry coefficient, γ , is statistically significant. Similarly, if we consider the SV model with leverage (this model can capture leverage, and hence asymmetry), replacing (5) with (see Danielsson, 1994):

$$h_t = \phi_0 + \phi_1 h_{t-1} + \delta_1 \varepsilon_{t-1} + \delta_2 \left|\varepsilon_{t-1}\right| + \eta_{t-1} \qquad (6)$$

Then tests of the coefficients δ_1 and δ_2 could be associated with the significance of both the size and sign of shocks.

Tests of significance associated with single parameters or of model features are linked to diagnostic procedures based on the likelihood function. In fact, model comparison could also consider testing nested and/or non-nested models. In general, when we compare models belonging to the same family (such as within GARCH or SV), these are typically nested comparisons. Therefore, the validity of parametric restrictions could be evaluated by significance tests or, more appropriately, by Likelihood Ratio (LR) or Lagrange Multipliers (LM) tests.

In order to present some simple examples, the GJR(1,1) model nests the simple GARCH(1,1) model under a zero restriction on the parameter driving the asymmetry; APARCH nests GARCH which is obtained fixing the power coefficient to 2; SV model with asymmetry nests the simpler SV model under a zero parametric restriction similar to that of GJR. In these cases, assuming correct specification of the model (particularly of the innovation density), LM and LR tests have the standard asymptotic properties, and the LM statistic can be evaluated when the analytic score is available (see Fiorentini et al. (1996) for an example).

For a comparison of models belonging to separate (or non-nested) families of hypotheses, such as GARCH versus SV, or EGARCH versus GARCH, non-nested tests are required. Ling and McAleer (2000) and McAleer et al. (2007) propose simple procedures to compare GARCH and GJR models against the EGARCH model. Denote by $\hat{\sigma}_t^2$ the estimates of time t variance obtained from a GJR model, and consider the following EGARCH specification:

$$\ln\left(v_t^2\right) = \omega + \alpha\left|\eta_{t-1}\right| + \gamma\eta_{t-1} + \beta\ln\left(v_{t-1}^2\right) + \delta\ln\left(\hat{\sigma}_t^2\right) \tag{7}$$

where v_t^2 are the EGARCH conditional variances and η_t are the EGARCH standardized residuals. The test of the EGARCH null hypothesis against the GJR alternative corresponds to testing $\delta = 0$. Similarly, the test with GJR as the null involves a test of $\delta = 0$ in the auxiliary regression:

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 I\left(\varepsilon_{t-1} < 0\right) + \beta\sigma_{t-1}^2 + \delta\hat{v}_t^2 \tag{8}$$

where $\hat{v}_t^2$ is the estimate of the time t variance obtained from an EGARCH model. The corresponding tests for GARCH against EGARCH can be obtained as special cases of those given above.

A different test for GARCH against EGARCH was proposed in Lee and Brorsen (1997). The authors suggested a test based on the likelihood of two competing non-nested models, based on the procedures developed in Cox (1961, 1962). The Cox test compares two parametric models by evaluating the difference between maximum likelihood values as a deviation from its expectation. Lee and Bronsen (1997) evaluate the test statistic by using Monte Carlo methods. However, it is not clear if the conditions underlying the Cox test are satisfied. In particular, the two likelihoods should belong to separate families, that is, for a given parameter choice, the null hypothesis cannot be arbitrarily closely approximated by the alternative. A further aspect that may affect the validity of the test of the EGARCH model as the null hypothesis is that the statistical properties of EGARCH are as yet not known.

The approach of Cox (1961, 1962) is also closely related to the comparison methods outlined in Kim et at. (1998), who suggest a likelihood ratio test for non-nested models by obtaining the sampling distribution of the test statistic through Monte Carlo methods. In this case, the tested non-nested models are GARCH and Stochastic Volatility, making the Cox test more appropriate.

The procedure outlined in Kim et al. (1998) involves the following steps:

(1) Estimate the GARCH and SV model parameters and evaluate the corresponding likelihoods, denoted by $L_{SV}\left(x;\hat{\theta}_{SV}\right)$ and $L_{GARCH}\left(x;\hat{\theta}_{GARCH}\right)$, respectively, where the circumflex denotes estimated parameters, evaluate the likelihood ratio statistics:

$$LR_{SV,GARCH} = 2\left[\log L_{SV}\left(x;\hat{\theta}_{SV}\right) - \log L_{GARCH}\left(x;\hat{\theta}_{GARCH}\right)\right]$$

$$LR_{GARCH,SV} = 2\left[\log L_{GARCH}\left(x;\hat{\theta}_{GARCH}\right) - \log\log L_{SV}\left(x;\hat{\theta}_{SV}\right)\right],$$

where the first model represents the null hypothesis, and the SV density is evaluated by simulation methods, following the procedure in Kim et al. (1998);

(2) Simulate M paths under the null, estimate both models on each path, and evaluate the M likelihood ratio statistics;

(3) Test the null hypothesis using a Monte Carlo test, determining the p-value of the empirical likelihood ratio statistics under the simulated density of the LR test statistic.

Note that, the LR test statistic is not constrained to be positive as the two models are non-nested. Moreover, by reversing the null and alternative hypotheses, the test outcomes may lead to rejection or non-rejection of both models as the respective null hypotheses. Clearly, the procedure outlined in Kim et al. (1998) derives an approximate LR statistic density, and is also influenced by the fact that the true parameters are not known. To state the obvious, this test is computationally intensive.

Kobayashi and Shi (2005) propose a closely related test for EGARCH against SV. Their approach differs from the previous method as they modify the SV model. In fact, they consider the following SV parameterization:

$$\varepsilon_t = \sigma_z \exp\left(\frac{1}{2} h_t\right) z_t \tag{9}$$

$$h_t = \phi_0 + \phi_1 h_{t-1} + \alpha \frac{\varepsilon_{t-1}}{\exp\left(h_{t-1}\right)} + \beta \left| \frac{\varepsilon_{t-1}}{\exp\left(h_{t-1}\right)} \right| + \sigma_\eta \eta_t \tag{10}$$

$$\begin{pmatrix} z_t \\ \eta_t \end{pmatrix} \square\ D\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \tag{11}$$

The model of Kobayashi and Shi (2005) is a slightly modified version of the model in Danielsson (1994), where the volatility equation includes a dependence on both the sign and size of the standardized innovations. Notably, the model includes both asymmetry and leverage as the parameters need not be positive.

In the context of the slightly modified SV model, EGARCH is associated with the parametric restriction, $\sigma_\eta^2 = 0$. Kobayashi and Shi (2005) propose a Lagrange Multiplier (LM) test for the null hypothesis $\sigma_\eta^2 = 0$ (EGARCH) against an alternative of positive variance for the volatility equation. The LM test has an advantage that only EGARCH needs to be estimated. The Monte Carlo simulations reported to verify the size and power of the test show that the LM test for EGARCH against SV has good size and reasonable power (but the results would seem to be heavily dependent on the values of the parameters).

In addition to hypothesis testing approaches, information criteria may also be considered to compare models by using their likelihood penalized by a function of the number of parameters and number of sample observations. These methods allow a comparison of models where the conditional variances depend on observable quantities, such as GARCH and EGARCH, but cannot be applied to compare GARCH and SV as the likelihood function for SV models differs from that of conditional variance specifications (see Kim et al. (1998) for an example of the evaluation of the SV likelihood by simulation methods).

Alternative models of variances and volatility may also be compared through their ability to capture the heteroskedasticity inherent in financial time series. The most common approach for diagnostic checking is the Ljung-Box test statistic applied to the squared standardized residuals, with the preferred model as the one that permits greater whitening of the residuals. Furthermore, distributional hypotheses could also be considered in order to evaluate which

density is closer to the analyzed data. Standard tests such as the Jarque-Bera normality test, or the more general Kolmogorov-Smirnov, may be considered in this context.

In-sample comparisons, and the subsequent choice of the best model, may be optimal for structural analysis, but it does not guarantee an optimal choice for out-of-sample forecasting. In this case, the literature provides a number of alternative approaches for model comparison. In the following section, we present some that are tailored for comparing conditional variance models.

2.2 Out-of-sample comparisons

A comparison of SV and GARCH models out-of-sample may follow two different approaches: a direct comparison of variance forecasts, or an indirect comparison of variance models through the possible uses of the corresponding variance forecasts. This dichotomy follows from Patton and Sheppard (2009), who present a number of alternative theoretical approaches.

2.2.1. Direct model evaluation

Within the direct comparison, alternative models are contrasted by tests directly based on variance forecasts. Denote by $\hat{\sigma}_{j,t}^2$ the time *t* variance forecast of model *j*, and by σ_t^2 the true and unknown variance at time *t*. For each model we may evaluate, over a given forecast horizon, a set of standard quantities. Two well known examples are the Mean Absolute Error (MAE) and Mean Squared Error (MSE):

$$MAE(j) = \frac{1}{m}\sum_{t=1}^{m}\left|\sigma_t^2 - \hat{\sigma}_{j,t}^2\right| \tag{12}$$

$$MSE(j) = \frac{1}{m}\sum_{t=1}^{m}\left(\sigma_t^2 - \hat{\sigma}_{j,t}^2\right)^2 . \tag{13}$$

Given these quantities for each model, the preferred model will typically have lower values of both MAD and MSE, meaning lower deviations from the true variance.

A closely related comparison method is the use of Mincer-Zarnowitz (1969) regressions, where the variance forecasts are used as explanatory variables for the true variance:

$$\sigma_t^2 = \alpha + \beta \hat{\sigma}_{j,t}^2 + \varepsilon_t \,. \tag{14}$$

In this alternative framework, optimal models should have $\alpha = 0$ and $\beta = 1$, with a higher value of R^2. Therefore, models providing appropriate or similar coefficient values in (14) could be ranked by means of R^2 values.

However, two problems arise in both the Mincer-Zarnowitz-type regressions and in the use of MSE or MAE: (i) the true variance is not known; and (ii) ranking models on the basis of one or more statistical indicators is not necessarily a formal statistical test.

With respect to the first issue, unbiased estimates of the true variance could be recovered by realized volatility estimators (see Barndorff-Nielsen and Shephard (2002a,b) and Barndorff-Nielsen et al. (2008), among others). When high-frequency data are not available, the true variance could be approximated by the squared de-meaned return observed at time *t*, at the cost of a large noise component. Nevertheless, in the case of the Mincer-Zarnowitz regressions, Meddahi (2002) shows that the rankings based on R^2 are consistent to the inclusion of noise in the proxy used for σ_t^2.

Model equivalence could be tested more formally, for instance, by the approach proposed by Diebold and Mariano (1995), and generalized by Patton (2010). We may compare models by using tests based on loss function differentials, whereas MAE and MSE could be considered as specific loss functions. As shown in Patton (2010), the use of proxies for the underlying true volatility induces distortions in the model ranking for some loss functions. Patton (2010) proves that two loss functions are robust to noisy volatility proxies, and allows an unbiased model ordering. These loss functions are the MSE and QLIKE, as given below:

$$MSE(j) = \frac{1}{m}\sum_{t=1}^{m}\left(h_t - \hat{\sigma}_{j,t}^2\right)^2 \tag{15}$$

$$QLIKE(j) = \frac{1}{m}\sum_{t=1}^{m}\left(\ln\hat{\sigma}_{j,t}^2 + \frac{h_t}{\hat{\sigma}_{j,t}^2}\right)^2 \tag{16}$$

where h_t is a proxy for the true unobserved volatility σ_t^2. Alternative models for σ_t^2 can be compared by tests of equal predictive ability, which are associated with the null hypothesis of the expected null loss function differential:

$$H_0 : E\left[MSE(j)\right] - E\left[MSE(i)\right] = E\left[MSE(j) - MSE(i)\right] = E\left[lf_{MSE,t}(j,i)\right] = 0 \tag{17}$$

where we may write a similar expression for QLIKE, and the expectation is evaluated using the sample counterparts reported in (15) and (16). Building on the results in Diebold and Mariano (1996), the test statistic is given as:

$$\tau_{MSE}(i,j) = \frac{LF_{MSE}(j,i)}{\sqrt{Var\left[lf_{MSE,t}(j,i)\right]}} \tag{18}$$

where

$LF_{MSE}(j,i) = \frac{1}{m}\sum_{t=1}^{m} lf_{MSE,t}(j,i) = \frac{1}{m}\sum_{t=1}^{m}\left(\left(h_t - \hat{\sigma}_{j,t}^2\right)^2 - \left(h_t - \hat{\sigma}_{i,t}^2\right)^2\right)$ and $Var\left[lf_{MSE,t}(j,i)\right]$ is a heteroskedasticity and autocorrelation consistent variance estimator (with identical equivalence relations available for the QLIKE loss function). The test statistic is asymptotically distributed as a standardized normal, which allows a simple evaluation of the null hypothesis. In fact, the test is equivalent to a significance test of the intercept in a regression of the loss function differentials $lf_{MSE,t}(j,i)$ over a constant, and is thus readily available in all computer software packages that implement robust linear regression methods.

A relevant limitation of the comparisons based on Diebold-Mariano type tests is that they represent pairwise comparisons, so that it is not possible to exclude a priori the possibility of having different model rankings associated with different robust loss functions. The literature contains several approaches that have attempted to resolve this issue, such as the Reality Check of White (2000), the Superior Predictive Ability test of Hansen (2005), and the Model Confidence Set (MCS) of Hansen et al. (2005).

We suggest the use of the Model Confidence Set as this method provides a confidence set of statistically equivalent models. The approach developed in Hansen et al. (2005) constitutes a testing framework for the null hypothesis of equivalence across models, which is described by mean of loss functions. By referring to the MSE loss function (similar quantities can be obtained for the QLIKE loss function), and assuming that the set $\mathcal{M}$ contains a number of different models used to produce forecasts in a given out-of-sample range, the null hypothesis of MCS is given as:

$$H_0 : E\left[lf_{MSE,t}\left(j,i\right)\right] = 0,\ i > j,\ \forall i, j \in \mathcal{M} \tag{19}$$

The null hypothesis can be tested by means of two test statistics proposed in Hansen et al. (2005), namely:

$$t_R = \max_{j,i\in\mathcal{M}} \left| \frac{LF_{MSE}\left(j,i\right)}{\sqrt{Var\left[lf_{MSE,t}\left(j,i\right)\right]}} \right| \tag{20}$$

$$t_{SQ} = \sum_{j,i\in\mathcal{M},j>i} \left(\frac{LF_{MSE}\left(j,i\right)}{\sqrt{Var\left[lf_{MSE,t}\left(j,i\right)\right]}} \right)^2 \quad . \tag{21}$$

Both tests statistics are based on a bootstrap estimate of the variance, $Var\left[lf_{MSE,t}\left(j,i\right)\right]$. As the distribution is non-standard, the rejection region is determined using bootstrap p-values under the null hypothesis. If the null of equal predictive ability across all models is rejected, the worst performing model is excluded from the set $\mathcal{M}$. Such a model is identified using:

$$j = \arg\max_{j\in\mathcal{M}} \left(\sum_{i\in\mathcal{M},i\neq j} LF_{MSE}\left(j,i\right) \right) \left(Var\left(\sum_{i\in\mathcal{M},i\neq j} LF_{MSE}\left(j,i\right) \right) \right)^{-1/2} \tag{22}$$

where the variance is again determined through bootstrap techniques. The equal predictive ability of the remaining models should also be tested, thereby iterating the evaluation of the test statistics in (20) and (21), and the identification of the worst performing model in (22).

The procedure stops when the null hypothesis of equal predictive ability of the models still included in the set is not rejected. Subsequently, the MCS method provides a set of statistically equivalent models with respect to a given loss function. It should be noted that the optimal model set could contain a single model.

2.2.2. Indirect model evaluation

Indirect evaluation methods consider the uses of alternative variance forecasts. For instance, conditional variances could be used to price derivatives, or to define the market risk exposure of a portfolio. The literature has recently addressed the topic, focusing mainly on multivariate models (for example, see Caporin and McAleer (2010), Clements et al. (2009), Patton and Sheppard (2009), and Laurent et al. (2009), among others). At the univariate level, the approaches are much more widespread and have generally focused on specific applications. Many studies dealt with the evaluation of alternative GARCH specifications within a Value-at-Risk (VaR) framework (for example, see Caporin (2008), Berkowitz (2001), and Lopez (1999, 2001)).

Considerable empirical research has focused on tests for the evaluation of VaR forecasts. These are used to determine if a model is more appropriate with respect to competitors in determining the future expected risk of a financial instrument (such as a financial portfolio). In this framework, consider a variable displaying heteroskedasticity, possibly characterized by a time-varying mean, and with an unspecified conditional density (with additional parameters contained in the vector θ):

$$x_t \mid I^{t-1} \sim f\left(x_t; \mu_t, \sigma_t^2, \theta\right). \tag{23}$$

The one-day VaR for x_t is defined as:

$$\alpha = \int_{-\infty}^{VaR(x_{t+1};\alpha)} f\left(x_{t+1}; E\left[\mu_{t+1} \mid I^t\right], E\left[\sigma_{t+1}^2 \mid I^t\right], \hat{\theta}\right) dx_{t+1} \tag{24}$$

where the time-varying mean and variance are replaced by their conditional expectations, the additional parameters are estimated, and α is the VaR confidence level. Under normality, the

VaR has a simpler expression, namely $VaR(x_{t+1};\alpha) = E\left[\mu_{t+1} \mid I^t\right] + \Phi^{-1}(\alpha)\sqrt{E\left[\sigma^2_{t+1} \mid I^t\right]}$, where $\Phi^{-1}(\alpha)$ is the α-quantile of the standardized normal. Thus, VaR depends on the models used to capture the mean and variance dynamics.

The evaluation of alternative mean and variance specifications by using VaR could follow two approaches: (i) test if the VaR out-of-sample forecasts satisfy the condition $E\left[I\left(x_t < VaR(x_t;\alpha)\right)\right] = \alpha$, that is, if the expected number of VaR violations (namely, where returns are lower than the forecast VaR) is equal to the VaR confidence level; (ii) compare models by means of loss functions. Tests include the traditional method of Kupiec (1995) which, as shown in Lopez (1999, 2001) and Caporin (2008), have limited power in discriminating across alternative variance specifications. Thus, loss functions should be preferred, making an indirect comparison of GARCH and SV models very similar to the direct comparison. In the following, we provide an interpretation of VaR model comparisons by means of the Model Confidence Set which, to the best of our knowledge, would seem to be novel.

Loss functions based on VaR forecasts have been proposed, for instance, by Lopez (1999) and Caporin (2008). We suggest the following:

i) $IF = I\left(x_t < VaR(x_t;\alpha)\right)$; (25)

ii) $PIF_t = I\left(x_t < VaR(x_t;\alpha)\right)\left(1 + \left(x_t - VaR(x_t;\alpha)\right)^2\right)$; (26)

iii) $AD_t = \left|\,|x_t| - |VaR(x_t;\alpha)|\,\right| g(x_t)$; (27)

iv) $SD_t = \left(|x_t| - |VaR(x_t;\alpha)|\right)^2 g(x_t)$; (28)

v) $ASD_t = AD_t + \lambda SD_t$; (29)

vi) $RL_t = \max\left(VaR(x_t;\alpha); \frac{p_t}{60}\sum_{j=1}^{60} VaR(x_{t-j};\alpha)\right)$. (30)

In the previous list, the first function (the indicator loss function, IF) identifies exceptions, while the second penalizes exceptions by using the squared deviation between realized returns and VaR (penalized indicator function – PIF). The third and fourth loss functions could be read as first-order and second-order losses, respectively, between VaR and realized returns (Absolute Deviation, AD, and Squared Deviation, SD, loss functions, respectively).

They both depend on $g(x_t)$, a function of the observed variable, x_t, that focuses the loss functions, for instance, only on negative returns $g(x_t) = I(x_t < 0)$, on VaR violations $g(x_t) = I(x_t < VaR(x_t;\alpha))$ or, finally, on the entire returns path (if set equal to 1). In the fifth loss function, we combine the previous two, adding a parameter, λ, to modify the weight of a component (which can be used to increase or decrease the impact of squared deviations). Note that the fourth loss function is equivalent to the second if $g(x_t) = I(x_t < 0)$ and $VaR(x_t;\alpha)$ is always negative. Finally, the last loss function is also known as Regulatory Loss, and depends on a penalty term, p_t, which is calibrated over the number of exceptions in the last 250 days (3 up to 4 exceptions, 3.4, 3.5, 3.65, 3.75, 3.85 for 5, 6, 7, 8, and 9 exceptions, respectively, and 4 for more than 9 exceptions).

One striking advantage of these loss functions is that they are not based on the true volatility, but still depend on the volatility forecasts. Thus, they could be used within a MCS framework to compare alternative models, without suffering from the problems associated with the replacement of the true variance by a noisy proxy.

These methods could also be used in the multivariate framework and be applied to portfolios, in which the included asset variances follow a heteroskedastic density.

3. Empirical Example

In this section we present an empirical comparison of the methods discussed above. Daily stock market total return indices, as reported in Table 1, are examined for 2000-2009. We consider the large cap stock market indices of France (CAC40), Germany (DAX), Switzerland (SMI), Hong Kong (HS), and USA (S&P500). Returns are computed from index levels as $r_t = 100\left[\ln(I_t) - \ln(I_{t-1})\right]$. For each series, we report the descriptive statistics and sample period, which differ across market indices as holidays have been removed from the data on a single series basis, and these are not common over the countries considered.

For each return series, we fit four specific models, namely GARCH(1,1), GJR(1,1), EGARCH(1,1), and SVOL(1). The models are estimated on a rolling basis, using a window of 1000 observations, and under normality. The four models are then used to produce one-step-ahead variance forecasts, from 1 January 2004.

The models are compared using some of the methods described in the previous section. In particular, we consider the Ling and McAleer (2000) test for comparing GJR and GARCH against EGARCH, the Likelihood Ratio test in comparing GARCH against GJR, the Diebold-Mariano test using the MSE and QLIKE loss functions across all model pairs, and the Model Confidence Set approach using the MSE and QLIKE loss functions, the loss functions in

(26)-(30) with three VaR levels (1%, 5% and 10%), and the loss function in (31) with the 1% VaR level. For the ASD loss function, we set $\lambda=1$. Furthermore, in order to verify the stability of results over time, we compare the models over different out-of-sample periods, and we consider annual comparisons from 2004 to 2009 (that is, for 5 different years).

We start with the in-sample comparison of models using the Ling and McAleer (2000) and LR tests. As we estimated the models over a rolling sample of 1000 observations, we have a set of around 1500 estimates of all models (for estimation samples ending from 31 December 2003 to 30 December 2009). The number of estimates is not equal across all series as these differ with respect to national holidays. Table 2 reports the percentage of rejections of the null hypothesis at the 5% confidence level over the entire set of estimates available for each series.

Table 2 highlights that GJR(1,1) is always preferred to its GARCH(1,1) counterpart for the CAC40, DAX, SMI, and S&P500 indices, while only for the HS index does GJR not improve in-sample over GARCH in 32% of cases.

A different picture emerges when we compare non-nested models, namely GARCH and GJR against EGARCH. We use the Ling and McAleer (2000) test and consider four possible comparisons, modifying the null and alternative models accordingly. The Ling and McAleer (2000) test adds the fitted variances under the alternative to the auxiliary regression equation for the conditional variance equation under the null. A significant coefficient of the added variable provides evidence against the null model. The results for DAX, SMI and S&P500 are quite similar in that there is a large fraction of rejections when the null model is the GARCH and GJR specification, and a small fraction of rejections when the null model is EGARCH. Therefore, EGARCH is the preferred conditional volatility model. This finding is not surprising as EGARCH is more flexible than GARCH and GJR, can exhibit asymmetry and leverage, and there are no restrictions on the parameters of the model.

However, for CAC40 and HS, the results do not support a particular model, either suggesting that any alternative model is an improvement over the null (CAC40) in a large fraction of cases, or that no model can improve the null (HS) (again in a large percentage of cases).

In order to shed some light on this result, we recomputed Table 2 over two subsamples, 2004-2006 and 2007-2009, and the outcome is reported in Table 3. We do not report the LR test as the outcomes are stable across the two subsamples, with the exception of the HS index (for this index, the rejection frequency is higher in the second subsample).

Table 3 suggests EGARCH is optimal for the S&P500 index over the period 2004-2006, while asymmetry is significant for the period 2007-2009, that is, the GARCH estimates clearly improve when we include the EGARCH variances, there is little to choose between GJR and EGARCH.

For SMI, the empirical results are contrary to the above. GARCH is clearly rejected for 2004-2006, but there is no clear preference between GJR and EGARCH. In 2007-2009, EGARCH

performs better as compared with GJR and GARCH. The results for the DAX and CAC40 indices are similar to those of SMI for 2007-2009, while for 2004-2006 there is no clear preference across the alternative models for DAX. The results for CAC40 suggest a mild preference for GJR.

Finally, for the HS index, the evidence suggests a small percentage of rejections of the null hypothesis, alluding to the fact thatt most models provide very similar conditional variance patterns.

Moving to the out-of-sample comparison, we start from the Diebold-Mariano test outcomes (direct evaluation method) using the MSE and QLIKE loss functions. In order to evaluate model performance across different market phases, we consider separately each out-of-sample year. Table 4 reports some salient empirical findings (the full set of empirical results is available from the authors upon request).

Focusing on the MSE loss function, all empirical models seem very similar for all stock market indices, with the null hypothesis of zero loss function differential being rejected only in few cases. When we consider the QLIKE loss function, the null hypothesis is rejected more frequently, with the finding seemingly independent of the sample used for model evaluation (the results are similar for 2006 and 2008, two years with very different volatility and returns). In this case, there are some differences across the stock market indices, but the outcomes suggest a preference for GJR over GARCH, and of GJR and EGARCH over SVOL. Furthermore, GJR and EGARCH are generally equivalent.

Although some preference ordering across models may appear in some cases, the limitation of the Diebold-Mariano test is that it only considers pairwise comparisons. As suggested in Section 2, the Model Confidence Set method overcomes this restrictive comparison.

A number of tables collects the results over the entire set of loss functions, and over the out-of-sample years and stock market indices. Tables 5-9 report the Model Confidence Set results based on the R statistic in (21) for selected out-of-sample periods. The results for the statistic SQ are equivalent and are not reported (the entire set of results is available from the authors upon request).

For each stock market index, we evaluate the four alternative models by using the MSE and QLIKE loss functions, as well as the loss functions defined in (26)-(30).

If we consider the S&P500 index (Table 5), the results differ across the out-of-sample evaluation periods. In 2004, all models are equivalent as they are all included in the confidence set independently of the loss function used for their evaluation. For 2006, some differences appear across the loss functions. For MSE and QLIKE, the optimal model is GJR; IF and PIF exclude, in most cases, SVOL from the confidence set; AD, SD, and ASD suggest that the optimal models are GJR and SVOL; finally, RL prefers the GARCH and EGARCH specifications. In summary, there is not a clear preference for a specific model. Model

preference depends on the loss function under consideration, and on the sample period used for model evaluation.

The last finding may be interpreted as confirmation of the in-sample and direct model comparison outcomes, which did not provide a clear indication of a single model. This interpretation is corroborated by the 2008 results for the S&P500 index: MSE considers all models as equivalent; QLIKE prefers GJR; IF, PIF and RL indicate a preference for GARCH and GJR; while AD, SD and ASD suggest that the optimal models are EGARCH and SVOL.

Similar patterns are observed for the other stock market indices in that all models are equivalent under some specific out-of-sample periods, and with model preferences changing with respect to the loss function used. Some behaviour is, however, common. When the Model Confidence Set includes fewer models than those that are available, the statistically equivalent models generally differ between the IF-PIF-RL loss functions and the AD-SD-ASD loss functions. The former indicate a preference for GARCH and GJR, while the latter tend to support EGARCH and SVOL.

Thus, it seems that the second set of loss functions has a preference for more flexible models. Such behaviour may depend on the structure of the loss functions themselves: AD and SD (and hence also ASD) monitor the entire evolution of conditional variances without focusing on the exceptions or without penalizing VaR with respect to past violations. The inevitable conclusion to be drawn is that when we give a large relevance to volatility spikes, most models appear relevant, and simple specifications may perform as well as their more flexible counterparts. If we consider the evolution over time of the conditional volatility, then more flexible models are to preferred.

4. Concluding Remarks

In this paper we reviewed some existing methods for model selection and testing of non-nested univariate volatility models. We first considered in-sample methods, such as nested and non-nested hypothesis testing, and diagnostic checking procedures (such as Ljung-Box and distributional hypotheses). We then focused on out-of-sample approaches based on model forecast evaluation. Starting from the traditional mean squared error and mean absolute error criteria, we considered more general loss functions based on Value-at-Risk forecasts, compared by means of the Model Confidence Set approach. Finally, we presented an empirical example using the less common approaches for model comparison, namely non-nested hypothesis testing and VaR-based loss functions.

The paper was based on simple univariate specifications focusing on volatility asymmetry and leverage. The proposed loss function approaches can easily be used on the forecasts produced by other univariate specifications, as well as multivariate models.

Table 1: Sample Statistics of Index Returns

Stock Market Index	Number of observations	Mean	Standard Deviation	Min	Max	Asym.	Kurt.
CAC40	2552	-0.016	1.577	-9.472	10.595	0.026	7.953
DAX	2542	-0.005	1.674	-7.433	10.797	0.072	7.094
SMI	2522	-0.004	1.313	-8.108	10.788	0.072	8.970
HS	2489	0.009	1.708	-13.582	13.407	-0.038	10.612
S&P500	2514	-0.011	1.401	-9.470	10.957	-0.104	10.662

Table 2: Rejection Percentages of the Null Hypothesis in the Full Sample

Test	Null model	Alternative	Market Index				
			CAC40	DAX	SMI	HS	S&P500
LR test	GARCH	GJR	100.00%	100.00%	100.00%	68.36%	99.80%
Ling-McAleer	GARCH	EGARCH	70.44%	65.25%	84.79%	26.41%	83.44%
Ling-McAleer	GJR	EGARCH	44.40%	59.24%	52.14%	9.98%	65.53%
Ling-McAleer	EGARCH	GARCH	63.60%	36.12%	23.30%	39.05%	32.58%
Ling-McAleer	EGARCH	GJR	55.66%	21.68%	22.84%	39.78%	39.47%

Note: The null hypothesis of the LR test is associated with a preference for the GARCH model against the GJR. For the Ling and McAleer (2000) test, the alternative model column denotes the model whose variances are used as additional explanatory variables in the dynamics governing the variances as given by the null model. The rejection of the null hypothesis is associated with a non-significant coefficient and signals a preference for the null model over the alternative one.

Table 3: Rejection Percentages of the Null Hypothesis in Two Subsamples

Test	Null model	Alternative	Market Index				
			CAC40	DAX	SMI	HS	S&P500
2004 to 2006							
Ling-McAleer	GARCH	EGARCH	47.99%	43.15%	75.39%	22.52%	75.63%
Ling-McAleer	GJR	EGARCH	10.89%	34.79%	13.55%	18.10%	61.46%
Ling-McAleer	EGARCH	GARCH	63.81%	13.95%	10.53%	42.36%	10.86%
Ling-McAleer	EGARCH	GJR	73.67%	8.74%	12.50%	43.30%	13.51%
2007 to 2009							
Ling-McAleer	GARCH	EGARCH	92.95%	87.66%	94.20%	30.25%	88.08%
Ling-McAleer	GJR	EGARCH	78.07%	83.99%	90.78%	1.98%	67.93%
Ling-McAleer	EGARCH	GARCH	63.32%	58.53%	36.10%	35.80%	53.43%
Ling-McAleer	EGARCH	GJR	37.47%	34.78%	33.20%	36.33%	64.49%

Note: In the Ling and McAleer (2000) test, the alternative model column denotes the model whose variances are used as additional explanatory variables in the dynamics governing the variances as given by the null model. The rejection of the null hypothesis is associated with a non-significant coefficient and signals a preference for the null model over the alternative one.

Table 4: Diebold-Mariano Test Statistics for Selected Years

Index	GJR GARCH	EGARCH GARCH	SVOL GARCH	EGARCH GJR	SVOL GJR	SVOL EGARCH
MSE loss function – out-of-sample period: 2004						
CAC40	-0.917	-0.926	1.008	-0.215	1.239	1.191
DAX	-0.206	-1.087	1.192	-1.334	1.085	1.848
SMI	1.726	0.934	1.496	-1.145	0.018	0.689
HS	-1.520	-1.610	0.363	-0.758	1.082	1.246
S&P500	-1.940	-0.658	-0.201	1.127	1.441	0.505
QLIKE loss function – out-of-sample period: 2004						
CAC40	**-2.181**	-0.252	0.705	0.825	1.791	0.507
DAX	**-3.057**	**-2.362**	1.158	-0.562	*3.038*	*2.840*
SMI	-0.112	0.032	*2.177*	0.308	1.686	1.555
HS	**-2.368**	-1.889	1.173	-0.283	*2.128*	*2.195*
S&P500	**-2.383**	-0.621	0.089	1.019	1.953	0.637
MSE loss function – out-of-sample period: 2006						
CAC40	-0.522	-0.610	1.309	0.150	1.005	1.136
DAX	-0.740	0.057	*2.215*	1.674	1.857	1.534
SMI	-0.463	-1.327	1.140	-0.307	0.977	1.254
HS	-1.005	-1.845	0.652	-1.562	0.740	1.025
S&P500	**-2.422**	1.464	0.743	1.614	1.921	-1.442
QLIKE loss function – out-of-sample period: 2006						
CAC40	-1.931	-1.489	*2.025*	0.866	*2.280*	*2.351*
DAX	**-2.043**	-0.178	*2.408*	1.505	*2.962*	*3.032*
SMI	**-2.717**	**-3.555**	1.551	0.670	1.799	1.867
HS	-0.898	-1.066	0.931	-0.906	0.979	1.080
S&P500	**-2.198**	0.445	1.049	*2.082*	*2.510*	0.422
MSE loss function – out-of-sample period: 2008						
CAC40	-1.758	**-2.082**	0.598	0.517	1.166	1.400
DAX	-1.898	-1.746	0.990	1.347	1.493	1.482
SMI	-1.404	-1.724	0.890	0.586	1.112	1.219
HS	-1.734	**-2.013**	0.806	-0.251	1.210	1.436
S&P500	-1.509	-1.046	1.826	0.574	1.892	1.844
QLIKE loss function – out-of-sample period: 2008						
CAC40	-1.699	-0.746	1.940	0.765	*2.020*	1.641
DAX	-1.786	-0.786	*2.570*	0.761	*2.549*	*2.355*
SMI	-1.916	-0.594	1.734	1.310	1.948	1.865
HS	**-3.017**	**-2.934**	1.963	-0.816	*2.015*	*2.048*
S&P500	**-2.398**	0.616	*3.246*	*2.036*	*3.346*	*3.025*

Note: The test evaluates the null of zero expected difference between the loss function of the first row model minus the loss function of the second row model. The test statistic is distributed as a standardized normal. Significant values (5% confidence level) indicate a preference for the first row model (if negative - in bold) or for the second row model (if positive - in italics underlined).

Table 5: S&P500 Model Confidence Set

	MSE	QLIKE	IF			PIF			AD			SD			ASD			RL
			1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	
Out-of-sample period: 2004																		
GARCH	**0.15**	0.04	**0.31**	**1.00**	**1.00**	**0.27**	**1.00**	**1.00**	**0.16**	**0.19**	**0.27**	**0.24**	**0.28**	**0.33**	**0.24**	**0.27**	**0.30**	**0.69**
GJR	**1.00**	**1.00**	**1.00**	**0.34**	**0.62**	**1.00**	**0.55**	**0.84**	**0.42**	**0.45**	**0.70**	**0.71**	**0.78**	**0.88**	**0.55**	**0.59**	**0.77**	**0.40**
EGARCH	**0.34**	**0.24**	**0.31**	**0.12**	**0.16**	**0.27**	**0.14**	**0.23**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.20**
SVOL	**0.34**	0.07	**0.31**	**0.12**	**0.25**	**0.27**	**0.14**	**0.40**	**0.78**	**0.76**	**0.70**	**0.97**	**0.94**	**0.88**	**0.88**	**0.84**	**0.77**	**1.00**
Out-of-sample period: 2006																		
GARCH	0.06	0.06	**0.17**	0.06	**0.18**	**0.57**	**1.00**	**0.21**	0.02	0.01	0.01	0.02	0.03	0.04	0.02	0.02	0.01	**0.29**
GJR	**1.00**	**1.00**	**0.17**	**1.00**	**0.74**	**0.26**	**0.33**	**1.00**	**0.32**	**0.34**	**0.59**	**0.31**	**0.37**	**0.60**	**0.30**	**0.36**	**0.58**	0.04
EGARCH	0.06	0.06	**1.00**	0.06	**1.00**	**1.00**	**0.33**	**0.94**	0.02	0.01	0.01	0.02	0.03	0.04	0.02	0.02	0.01	**1.00**
SVOL	0.06	0.06	**0.17**	0.04	0.06	**0.26**	0.08	0.03	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.01
Out-of-sample period: 2008																		
GARCH	**0.27**	0.04	**0.46**	**0.56**	**1.00**	**0.43**	**0.48**	**1.00**	0.04	0.02	0.05	0.08	0.09	0.08	0.07	0.07	0.08	**0.63**
GJR	**1.00**	**1.00**	**1.00**	**1.00**	**0.47**	**1.00**	**1.00**	**0.37**	0.02	0.01	0.05	0.08	0.09	0.08	0.06	0.07	0.08	**1.00**
EGARCH	**0.52**	0.04	0.07	0.04	0.01	0.06	0.03	**0.37**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.01
SVOL	**0.22**	0.04	0.07	0.01	0.01	0.06	0.03	**0.12**	**0.73**	**0.41**	**0.26**	**0.47**	**0.27**	**0.11**	**0.54**	**0.39**	**0.15**	0.04

Note: The table reports the Model Confidence Set over different loss functions and periods. Bold values denote the models that are included at the 10% confidence level in the confidence set (these models are statistically equivalent if compared using the loss function reported in the first and second rows). The loss functions names correspond to those in (16)-(17) and (26)-(31), the second row reports the Value-at-Risk confidence level (when needed).

Table 6: CAC40 Model Confidence Set

	MSE	QLIKE	IF			PIF			AD			SD			ASD			RL
			1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	
Out-of-sample period: 2004																		
GARCH	**0.60**	**0.11**	**0.44**	0.08	**0.12**	**0.87**	**0.76**	**0.21**	0.02	0.02	0.02	0.03	0.04	0.04	0.03	0.02	0.04	**0.30**
GJR	**0.79**	**1.00**	**1.00**	0.08	**0.12**	**1.00**	**0.25**	**0.18**	0.02	0.02	0.03	0.04	0.04	0.04	0.03	0.02	0.04	**0.30**
EGARCH	**1.00**	**0.45**	**0.44**	0.08	0.05	**0.58**	**0.13**	0.08	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.05
SVOL	**0.60**	**0.11**	**0.44**	**1.00**	**1.00**	**0.87**	**1.00**	**1.00**	0.02	0.02	0.03	0.04	0.04	0.04	0.03	0.02	0.04	**1.00**
Out-of-sample period: 2006																		
GARCH	**0.68**	0.08	**1.00**	**0.63**	**0.79**	**1.00**	**0.68**	**0.94**	**0.11**	**0.31**	**0.12**	**0.24**	**0.28**	**0.43**	**0.15**	**0.30**	**0.24**	**0.78**
GJR	**1.00**	**1.00**	0.06	**1.00**	**1.00**	**0.49**	**1.00**	**1.00**	**0.18**	**0.31**	**0.12**	**0.24**	**0.28**	**0.35**	**0.15**	**0.30**	**0.24**	**1.00**
EGARCH	**0.83**	**0.34**	0.06	**0.49**	**0.70**	**0.40**	**0.37**	**0.94**	**0.20**	**0.47**	**0.92**	**0.24**	**0.31**	**0.53**	**0.22**	**0.38**	**0.70**	**0.20**
SVOL	**0.38**	0.06	0.02	**0.49**	**0.40**	0.04	**0.37**	**0.44**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.10**
Out-of-sample period: 2008																		
GARCH	0.08	**0.11**	**1.00**	**0.22**	**0.43**	**1.00**	**0.40**	**1.00**	0.01	0.01	0.02	0.04	0.04	0.03	0.04	0.03	0.04	**1.00**
GJR	**1.00**	**1.00**	0.05	**1.00**	**1.00**	**0.19**	**1.00**	**0.84**	0.01	0.02	0.02	0.05	0.05	0.05	0.06	0.05	0.04	**0.46**
EGARCH	**0.45**	**0.42**	0.01	**0.13**	0.08	0.01	0.08	**0.73**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.02
SVOL	**0.14**	**0.11**	0.05	**0.13**	0.04	0.09	0.08	**0.61**	**0.85**	**0.58**	**0.29**	**0.57**	**0.24**	0.05	**0.65**	**0.32**	**0.11**	0.02

Note: The table reports the Model Confidence Set over different loss functions and periods. Bold values denote the models that are included at the 10% confidence level in the confidence set (these models are statistically equivalent if compared using the loss function reported in the first and second rows). The loss functions names correspond to those in (16)-(17) and (26)-(31), the second row reports the Value-at-Risk confidence level (when needed).

Table 7: DAX Model Confidence Set

	MSE	QLIKE	IF			PIF			AD			SD			ASD			RL
			1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	
Out-of-sample period: 2004																		
GARCH	**0.26**	0.00	**0.70**	**1.00**	**1.00**	**0.90**	**1.00**	**1.00**	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	**0.93**
GJR	**0.26**	**0.54**	**0.30**	**0.55**	**0.22**	**0.81**	**0.64**	**0.11**	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	**0.93**
EGARCH	**1.00**	**1.00**	**0.70**	0.02	**0.22**	**0.90**	0.01	**0.27**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.05
SVOL	**0.26**	0.00	**1.00**	**0.55**	**0.22**	**1.00**	**0.64**	**0.11**	**0.36**	**0.30**	**0.17**	**0.55**	**0.49**	**0.37**	**0.51**	**0.40**	**0.26**	**1.00**
Out-of-sample period: 2006																		
GARCH	**0.43**	0.08	0.06	**0.39**	**0.80**	**0.48**	**0.38**	**0.95**	0.04	**0.20**	**0.60**	**0.18**	**0.35**	**0.70**	**0.10**	**0.25**	**0.65**	**0.65**
GJR	**1.00**	**1.00**	0.06	**1.00**	**0.43**	**1.00**	**1.00**	**0.43**	0.04	**0.20**	**0.68**	**0.18**	**0.35**	**0.70**	**0.10**	**0.25**	**0.66**	**1.00**
EGARCH	**0.13**	0.08	**1.00**	**0.39**	**1.00**	**0.17**	**0.38**	**1.00**	0.04	**0.20**	**0.68**	**0.18**	**0.35**	**0.70**	**0.10**	**0.25**	**0.66**	**0.39**
SVOL	**0.12**	0.01	0.06	**0.10**	**0.10**	**0.17**	**0.10**	**0.14**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.03**
Out-of-sample period: 2008																		
GARCH	**0.10**	**0.10**	**0.69**	0.02	**1.00**	**0.70**	**0.13**	**0.74**	0.00	0.01	0.01	0.02	0.04	0.04	0.01	0.01	0.04	**1.00**
GJR	**1.00**	**1.00**	**1.00**	**1.00**	**0.69**	**1.00**	**1.00**	**1.00**	0.00	0.01	0.01	0.02	0.03	0.04	0.01	0.01	0.04	**0.40**
EGARCH	**0.14**	**0.54**	**0.43**	0.01	0.04	**0.26**	0.02	**0.46**	0.03	**0.13**	**0.71**	0.02	**0.42**	**1.00**	0.01	**0.26**	**1.00**	0.02
SVOL	**0.10**	0.05	0.01	0.00	0.03	0.05	0.02	**0.34**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.33**	**1.00**	**1.00**	**0.49**	0.01

Note: The table reports the Model Confidence Set over different loss functions and periods. Bold values denote the models that are included at the 10% confidence level in the confidence set (these models are statistically equivalent if compared using the loss function reported in the first and second rows). The loss functions names correspond to those in (16)-(17) and (26)-(31), the second row reports the Value-at-Risk confidence level (when needed).

Table 8: SMI Model Confidence Set

	MSE	QLIKE	IF			PIF			AD			SD			ASD			RL
			1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	
Out-of-sample period: 2004																		
GARCH	**1.00**	**0.95**	**0.14**	**1.00**	**0.80**	**0.22**	**1.00**	**0.82**	**0.69**	**0.69**	**0.81**	**0.98**	**0.98**	**0.95**	**0.93**	**0.91**	**0.81**	**0.94**
GJR	**0.43**	**1.00**	**0.14**	**0.44**	**1.00**	**1.00**	**0.46**	**1.00**	**0.69**	**0.69**	**0.81**	**0.71**	**0.71**	**0.79**	**0.67**	**0.70**	**0.79**	**0.46**
EGARCH	**0.49**	**0.95**	**1.00**	**0.66**	**0.51**	**0.22**	**0.61**	**0.49**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.09**
SVOL	**0.43**	0.06	**0.14**	**0.66**	**0.51**	**0.22**	**0.58**	**0.49**	**0.78**	**0.74**	**0.81**	**0.98**	**0.98**	**0.95**	**0.93**	**0.91**	**0.79**	**1.00**
Out-of-sample period: 2006																		
GARCH	**0.31**	0.00	**0.20**	**0.41**	**1.00**	**0.19**	**0.39**	**1.00**	0.03	0.01	0.02	0.07	0.06	**0.18**	0.05	0.03	0.05	**0.71**
GJR	**0.78**	**1.00**	**1.00**	**1.00**	**0.33**	**1.00**	**1.00**	**0.37**	**0.12**	**0.15**	**0.24**	**0.15**	**0.19**	**0.25**	**0.16**	**0.17**	**0.23**	**1.00**
EGARCH	**1.00**	**0.57**	**0.20**	**0.41**	**0.33**	**0.20**	**0.39**	**0.37**	**0.12**	**0.16**	**0.26**	**0.15**	**0.19**	**0.39**	**0.16**	**0.17**	**0.30**	**0.22**
SVOL	**0.31**	0.00	**0.18**	**0.25**	**0.33**	**0.16**	**0.22**	**0.29**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.05**
Out-of-sample period: 2008																		
GARCH	**0.14**	**0.14**	**0.28**	**0.86**	**0.93**	**0.42**	**1.00**	**1.00**	0.02	0.02	0.01	0.09	0.04	0.06	0.06	0.04	0.03	**1.00**
GJR	**1.00**	**1.00**	**1.00**	**0.46**	**1.00**	**1.00**	**0.68**	**0.98**	0.02	0.02	0.01	0.08	0.04	0.06	0.06	0.04	0.03	**0.64**
EGARCH	**0.52**	**0.15**	**0.15**	**0.18**	**0.13**	0.06	**0.18**	**0.29**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.03
SVOL	**0.21**	**0.14**	**0.15**	**1.00**	**0.93**	0.06	**0.60**	**0.79**	**0.19**	0.02	0.01	**0.24**	0.04	0.06	**0.22**	0.04	0.03	**0.48**

Note: The table reports the Model Confidence Set over different loss functions and periods. Bold values denote the models that are included at the 10% confidence level in the confidence set (these models are statistically equivalent if compared using the loss function reported in the first and second rows). The loss functions names correspond to those in (16)-(17) and (26)-(31), the second row reports the Value-at-Risk confidence level (when needed).

Table 9: HS Model Confidence Set

	MSE	QLIKE	IF			PIF			AD			SD			ASD			RL
			1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	
Out-of-sample period: 2004																		
GARCH	**0.36**	0.07	**0.69**	**0.71**	**1.00**	**0.74**	**0.77**	**1.00**	**0.28**	**0.47**	**0.46**	**0.39**	**0.43**	**0.42**	**0.31**	**0.43**	**0.46**	**1.00**
GJR	**0.45**	**0.76**	**0.31**	**0.83**	**0.43**	**0.70**	**1.00**	**0.13**	**0.28**	**0.47**	**0.46**	**0.39**	**0.43**	**0.42**	**0.31**	**0.43**	**0.46**	**0.55**
EGARCH	**1.00**	**1.00**	**0.31**	**0.71**	**0.43**	**0.70**	**0.67**	**0.57**	**0.86**	**1.00**	**1.00**	**0.74**	**0.81**	**0.93**	**0.77**	**0.95**	**1.00**	**0.18**
SVOL	**0.41**	0.07	**1.00**	**1.00**	**0.43**	**1.00**	**0.81**	**0.13**	**1.00**	**0.85**	**0.82**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.92**	**0.10**
Out-of-sample period: 2006																		
GARCH	**0.27**	**0.70**	**0.16**	**0.67**	0.01	**0.57**	**0.70**	**0.68**	0.01	0.01	0.02	0.03	0.03	0.02	0.01	0.01	0.03	**0.46**
GJR	**0.27**	**0.70**	**0.16**	**1.00**	0.01	**0.57**	**1.00**	**0.68**	0.01	0.01	0.02	0.03	0.03	0.02	0.01	0.01	0.03	**1.00**
EGARCH	**1.00**	**1.00**	**1.00**	**0.67**	**1.00**	**1.00**	**0.70**	**1.00**	0.01	0.00	0.02	0.03	0.02	0.01	0.01	0.01	0.01	**0.74**
SVOL	**0.27**	**0.70**	**0.16**	**0.67**	0.01	**0.38**	**0.70**	0.02	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.02
Out-of-sample period: 2008																		
GARCH	**0.24**	0.01	**0.31**	**1.00**	**0.81**	**0.13**	**1.00**	**0.26**	0.00	0.00	0.01	0.04	0.04	0.03	0.01	0.02	0.02	**1.00**
GJR	**0.73**	**0.36**	**1.00**	**0.19**	**1.00**	**1.00**	**0.93**	**1.00**	0.00	0.00	0.01	0.04	0.04	0.07	0.02	0.02	0.05	**0.20**
EGARCH	**1.00**	**1.00**	**0.31**	**0.19**	**0.81**	**0.24**	**0.57**	**0.40**	0.00	0.00	0.01	0.04	0.04	**0.63**	0.02	0.02	**0.36**	0.03
SVOL	**0.24**	0.01	0.04	0.00	0.00	**0.13**	0.03	**0.10**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.00

Note: The table reports the Model Confidence Set over different loss functions and periods. Bold values denote the models that are included at the 10% confidence level in the confidence set (these models are statistically equivalent if compared using the loss function reported in the first and second rows). The loss functions names correspond to those in (16)-(17) and (26)-(31), the second row reports the Value-at-Risk confidence level (when needed).

References

Asai, M., McAleer, M., 2005, Dynamic asymmetric leverage in stochastic volatility models, Econometric Reviews, 24, 317–332.

Asai, M., McAleer, M. and Yu, J., 2006, Multivariate stochastic volatility: a review, Econometric Reviews, 25(2-3), 145-175.

Barndorff-Nielsen, O.E. and Shephard, N., 2002a, Econometric analysis of realized volatility and its use in estimating stochastic volatility models, Journal of the Royal Statistical Society B, 64, 253-280.

Barndorff-Nielsen, O.E. and Shephard, N., 2002b, Estimating quadratic variation using realized variance, Journal of Applied Econometrics, 17, 457-477.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. and Shephard, N., 2008, Designing realized kernels to measure the ex post variation of equity prices in the presence of noise, Econometrica, 76, 1481-1536.

Berkowitz, J., 2001, The accuracy of density forecasts in risk management, Journal of Business and Economic Statistics, 19, 465-474.

Bollerslev T., 1986, Generalized autoregressive conditional heteroskedasticity, Journal of Econometrics, 31, 307-327.

Bollerslev, T., Chou, R.Y., and Kroner, K.F., 1992, ARCH modeling in finance: a review of the theory and empirical evidence, Journal of Econometrics, 52:,5-59.

Bollerslev T., Engle, R.F., and Nelson D.B., 1994, ARCH models. In Handbook of Econometrics, Engle, R.and McFadden, D. (eds.), North Holland, Amsterdam.

Bougerol, P. and Picard, N.M., 1992, Stationarity of GARCH processes and of some nonnegative time series, Journal of Econometrics, 52, 115–127.

Boussama, F., 2000, Asymptotic normality for the quasi-maximum likelihood estimator of a GARCH model, Comptes Rendus de l'Acad´emie des Sciences de Paris, Série I 331, 81-84.

Caporin, M., 2008, Evaluating value-at-risk measures in presence of long memory conditional volatility, Journal of Risk, 10-3, 79-110.

Caporin, M., and McAleer, M., 2009, Do we really need both BEKK and DCC? A tale of two covariance models. Available at SSRN: http://ssrn.com/abstract=1338190.

Caporin, M., and McAleer, M., 2010, Ranking multivariate GARCH models by problem dimension. Available at SSRN: http://ssrn.com/abstract=1601236.

Clements, A., Doolan, M., Hurn, S., and Becker, M., 2009, On the efficacy of techniques for evaluating multivariate volatility forecasts, NCER Working Paper.

Cox, D.R., 1961, Tests of separate families of hypotheses, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1, 105-123.

Cox, D.R., 1962, Further results on tests of separate families of hypotheses, Journal of the Royal Statistical Society B, 24, 406-424.

Danielsson, J., 1994, Stochastic volatility in asset prices: estimation with simulated maximum likelihood, Journal of Econometrics, 61, 375-400.

Diebold, F.X. and Mariano, R.S., 1995, Comparing predictive accuracy, Journal of Business and Economic Statistics, 13-3, 253-263.

Elie, L., and Jeantheau, T., 1995, Consistance dans les modèles hétéroscedastiques, Comptes Rendus de l'Académie des Sciences, Série I 320, 1255-1258.

Engle, R.F., 1982, Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation, Econometrica, 50, 987-1007.

Fiorentini, G., Calzolari, G., and Panattoni, L., 1996, Analytic derivatives and the computation of GARCH estimates, Journal of Applied Econometrics, 11, 399-417.

Glosten, L.R., Jagannathan, R., and Runkle, D.E., 1993, On the relation between expected value and the volatility of the nominal excess return on stocks, Journal of Finance, 8, 1779-1801.

Ghysels, E., Harvey, A.C., and Renault, E., 1996, Stochastic volatility. In Rao, C.R., and Maddala, G.S. (eds.), Statistical Models in Finance, Amsterdam, North-Holland, pp. 119-191.

Hansen, P.R., 2005, A test for superior predictive ability, Journal of Business and Economic Statistics, 23-4, 365-380.

Hansen, P.R., Lunde, A. and Nason, J.M., 2005, Model confidence sets for forecasting models, Federal Reserve Bank of Atlanta WP 2005-7.

Kobayashi, M., and Shi, X., 2005, Testing for EGARCH against stochastic volatility models, Journal of Time Series Analysis, 26(1), 135-150.

Kupiec, P., 1995, Techniques for verifying the accuracy of value-at-risk measurement models, Journal of Derivatives, 3, 73-84.

Kim, S., Shephard, N., and Siddhartha, C., 1998. Stochastic volatility: likelihood inference and comparison with ARCH models, Review of Economic Studies, 65, 361-393.

Laurent, S., Rombouts, J.V.K., and Violante, F., 2009, On loss functions and ranking forecasting performances of multivariate GARCH models, CIRANO WP 2009s-45.

Lee, J.H., and Brorsen, B.W., 1997, A non-nested test of GARCH vs. EGARCH models, Applied Economics Letters, 4, 765-768.

Ling, S., and McAleer, M., 2000, Testing GARCH versus E-GARCH, in Chan, W.-S., Li, W.K., and Tong, H. (eds.), Statistics and Finance: An Interface. Imperial College Press, London, pp. 226-242.

Ling, S., and McAleer, M., 2002a, Stationarity and the existence of moments of a family of GARCH processes, Journal of Econometrics 106, 109-117.

Ling, S., and McAleer, M., 2002b, Necessary and sufficient moment conditions for the GARCH(r,s) and asymmetric power GARCH(r,s) models, Econometric Theory, 18, 722-729.

Lopez, J.A., 1999, Regulatory evaluation of value-at-risk models, Journal of Risk, 1, 37-64.

Lopez, J.A., 2001, Evaluating the predictive accuracy of volatility models, Journal of Forecasting, 20, 87-109.

McAleer, M., 2005, Automated inference and learning in modeling financial volatility, Econometric Theory, 21, 232-261.

McAleer, M., Chan, F., and Marinova, D., 2007, An econometric analysis of asymmetric volatility: theory and applications to patents, Journal of Econometrics, 139, 259-284.

Meddahi, N., 2002, A theoretical comparison between integrated and realized volatilities, manuscript, Journal of Applied Econometrics, 17-5, 479-508.

Mincer, J., and Zarnowitz, V., 1969, The evaluation of economic forecasts, in Mincer, J. (ed.), Economic Forecasts and Expectations, Columbia University Press, New York.

Nelson, D.B., 1990, Stationarity and persistence in the GARCH (1,1) model, Econometric Theory, 6, 318–334.

Patton, A.J., 2010, Volatility forecast comparison using imperfect volatility proxies, Journal of Econometrics, forthcoming.

Patton, A.J., and Sheppard, K., 2009, Evaluating volatility and correlation forecasts, in Andersen, T.G., Davis, R.A., Kreiß, J.P., and Mikosch, T., (eds.), Handbook of Financial Time Series, Springer.

Shephard, N., 1996, Statistical aspects of ARCH and stochastic volatility. In: Cox, D. R., Hinkley, D. V., Barndorff-Nielsen, O. E. (eds.), Time Series Models in Econometrics, Finance and Other Fields, London, Chapman & Hall, pp. 1-67.

So, M.K.P., Li, W.K., and Lam, K., 2002, A threshold stochastic volatility model, Journal of Forecasting, 21, 473-500.

White, H., 2000, A reality check for data snooping, Econometrica, 68-5, 1097-1126.

Yu, J., 2005, On leverage in a stochastic volatility model, Journal of Econometrics, 127, 165-178.