Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Econometrics 131 (2006) 97–121

JOURNAL OF Econometrics

www.elsevier.com/locate/econbase

Consistent ranking of volatility models

Peter Reinhard Hansen[a,*], Asger Lunde[b]

[a] *Department of Economics, Stanford University, Landau Economics Building, 579 Serra Mall, Stanford, CA 94305-6072, USA*
[b] *Department of Marketing, Informatics and Statistics, Aarhus School of Business, Fuglesangs Allé 4, 8210 Aarhus V, Denmark*

Received 1 January 2004; accepted 1 April 2004
Available online 5 March 2005

Abstract

We show that the empirical ranking of volatility models can be inconsistent for the true ranking if the evaluation is based on a proxy for the population measure of volatility. For example, the substitution of a squared return for the conditional variance in the evaluation of ARCH-type models can result in an inferior model being chosen as 'best' with a probability that converges to one as the sample size increases. We document the practical relevance of this problem in an empirical application and by simulation experiments. Our results provide an additional argument for using the *realized variance* in out-of-sample evaluations rather than the squared return. We derive the theoretical results in a general framework that is not specific to the comparison of volatility models. Similar problems can arise in comparisons of forecasting models whenever the predicted variable is a latent variable.

JEL classification: C22; C52; C53

Keywords: Consistent ranking; Model comparison; Volatility models

*Corresponding author. Fax: + 1 650 725 5702.
E-mail addresses: peter.hansen@stanford.edu (P.R. Hansen), alunde@asb.dk (A. Lunde).

0304-4076/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2005.01.005

1. Introduction

The fact that population measures of volatility—such as the conditional variance and the integrated variance—are latent, makes it difficult to evaluate the performance of volatility models. A common resolution to this problem is to substitute a proxy for the true volatility and evaluate the models by comparing their predicted volatility to the proxy. For example, a squared return is often used as the dependent variable (as a proxy for the conditional variance) in the popular Mincer–Zarnowitz regression, see Poon and Granger (2003) for many references. Hansen and Lunde (2005a) contains an extensive out-of-sample evaluation of volatility models using six different loss functions. For comparisons that include both stochastic volatility models and GARCH models, see e.g. Kim et al. (1998) and Hansen et al. (2003), and for comparisons of simple ARMA models for the realized variance to GARCH models and other models, see Andersen et al. (2003) and Andersen et al. (2004).

The objective of this paper is to analyze how the substitution of a proxy for a latent variable can influence the ranking of competing volatility models. First we show that the substitution can induce a ranking of volatility models than is different from the intended ranking. This result has important implications for applied work, because the empirical ranking of volatility models can mislead practitioners to conclude that an inferior model is the 'best'. Second, we derive conditions that ensure the equivalence of the rankings that are induced by the true volatility and the proxy, respectively. We show that some of the popular criteria for evaluating and comparing volatility models do not satisfy the necessary conditions. Third, the practical relevance of our theoretical results are documented by three simulation experiments and an empirical comparison of volatility models that is based on IBM stock returns. Finally, our results provide an additional argument for using the realized variance (RV) as a proxy for the true volatility. The reason is that the RV is typically a more precise measure of volatility than is the squared return, so the former is less likely to induce an inconsistent ranking of volatility models.

To fix ideas, let σ_t^2 denote the population measure of volatility and let $\tilde{\sigma}_t^2$ denoted the corresponding proxy for σ_t^2, $t = 1, \ldots, n$. Further, let h_t be generic for a model-based measure of volatility and suppose that the 'precision' of h_t is assessed in terms of the expected loss, $E[L(\sigma_t^2, h_t)]$, where $L(\cdot, \cdot)$ is some known loss function. It is well-known that the substitution of $\tilde{\sigma}_t^2$ for σ_t^2 leads to an expected loss, $E[L(\tilde{\sigma}_t^2, h_t)]$, that need not be equal to $E[L(\sigma_t^2, h_t)]$ (this will typically follow by the Jensen inequality). So a criterion that is based on the proxy need not be useful for a *quantitative* assessment of volatility models. This point was forcefully made by Andersen and Bollerslev (1998) who showed that the substitution of squared returns for σ_t^2, (i.e. $\tilde{\sigma}_t^2 = r_t^2$), will spuriously indicate that ARCH-type models perform poorly. The focus of the present paper is the *qualitative* assessment of volatility models. Thus we are interested in the ranking of volatility models and we seek conditions that will ensure that $E[L(\sigma_t^2, h_{i,t})] < E[L(\sigma_t^2, h_{j,t})]$ if and only if $E[L(\tilde{\sigma}_t^2, h_{i,t})] < E[L(\tilde{\sigma}_t^2, h_{j,t})]$, where the subscripts, i and j, refer to two competing volatility models. Both $E[L(\sigma_t^2, \cdot)]$ and $E[L(\tilde{\sigma}_t^2, \cdot)]$ induce preorderings (rankings) of an arbitrary set of volatility models.[1] Similarly, the sample average, $n^{-1}\sum_{t=1}^{n} L(\tilde{\sigma}_t^2, \cdot)$, induces a third preordering and we refer to the three preorderings as the *true preordering*, the *approximate preordering*, and the *empirical preordering*, respectively.

We are interested in the equivalence of these preorderings. We shall refer to the discrepancy between the empirical preordering and the approximate preordering as the *sampling error*. This error will typically vanish asymptotically because $n^{-1}\sum_{t=1}^{n} L(\tilde{\sigma}_t^2, \cdot) \overset{p}{\to} E[L(\tilde{\sigma}_t^2, \cdot)]$ under standard regularity conditions, but in finite samples the sampling error makes it more difficult to identify the best volatility models. The potential discrepancy between the true preordering and the approximate preordering is more severe and we refer to this as the *objective-bias*.[2] Since the objective-bias is entirely due to (the distribution of) the measurement error, $\tilde{\sigma}_t^2 - \sigma_t^2$, this aspect is highly unfortunate for obvious reasons.

We show that some (but not all) of the commonly used criteria for comparing volatility models satisfy the conditions that ensure that the true preordering and the approximate preordering are equivalent. One important exception is the R^2-criterion, when the R^2 is calculated from the Mincer–Zarnowitz regression using logarithmically transformed variables. So this criterion may identify an inferior volatility model as the ‘best’, and the inferior model may spuriously be found to be ‘significantly’ better than all other models, with a probability that converges to one as the sample size increases.

We derive the theoretical results in a general framework that is not specific to the comparison of volatility models, and our results are also relevant for comparisons and the selection of forecasting models, whenever the predicted variable is unobserved.[3] More specifically, our theoretical results is applicable to comparisons of forecasting models whenever the variable that is being forecasted is a latent variable. For example, conditional quantile forecasts and conditional density forecasts belong to this category. See Giacomini and Komunjer (2005) and Corradi and Swanson (2005) for ways to evaluate and compare conditional forecasts of this type. We can view the predicted variable as an unknown ‘parameter’ in the loss function, which allows us to rephrased the problem as uncertainty about the loss function. This is related to Elliott et al. (2005) who consider the problem of estimating unknown loss function parameters.

The objective-bias should not be confused with the issues that can arise when estimation and evaluation of models is undertaken with different criteria functions,

[1] Since we cannot rule out the case $E[L(\sigma_t^2, h_{i,t})] = E[L(\sigma_t^2, h_{j,t})]$ for some $i \neq j$, ‘preordering’ is the correct terminology.

[2] The standard integrated variance (IV) regression model provides a good analogy. The least squares estimator, $\hat{\beta}_{\text{LS}}$, is consistent for some value, β^* say. However, β^* may not be the parameter of interest, β^0 say, which is the probability limit of the IV estimator. The ‘sampling error’ corresponds to the discrepancy between $\hat{\beta}_{\text{LS}}$ and β^* that typically vanishes as the sample size increases. The ‘objective-bias’ corresponds to the discrepancy between β^* and β^0.

[3] For comparisons and selection of forecasting models, see Diebold and Mariano (1995), West (1996), Sin and White (1996), White (2000), Hansen (2001, 2003), Clark and McCracken (2001), Corradi et al. (2001), Swanson and Zeng (2001), Perez-Amaral et al. (2003), Inoue and Kilian (2005), Giacomini and White (2003), and Hansen et al. (2004).

e.g., estimation by maximum likelihood and evaluation by some (other) loss function, see e.g. Skouras (2001). Nor should it be confused with the effects that the estimation of model-parameters can have on the evaluation, see e.g. Rossi (2005).

It is well known that the RV is a useful empirical measure of the IV and the conditional variance, see Barndorff-Nielsen and Shephard (2002a) and Meddahi (2002). See also Andersen et al. (2003) and references therein. In the context of evaluating volatility models, the RV can greatly reduce the sampling error which makes the evaluation 'more precise'. Another advantage of using the RV in this context is that this measure can eliminate the objective-bias. These two properties follow from the fact that the RV is a more precise measure of the IV than is the squared return. The gains from using the RV (rather than a squared return) are evident from our empirical out-of-sample comparison of volatility models.

This paper is organized as follows. We present the general theoretical framework in Section 2 and apply the results to the comparison of volatility models in Section 3. Section 4 contains empirical and simulation-based comparisons of volatility models, which document that the objective-bias has practical relevance. Section 5 contains a summary of our results and some concluding remarks. The appendix describes the Fourier method for calculating the RV-measure and all proofs.

2. The general theoretical framework

Let $\mathscr{X}$ be a random variable that is evaluated through the expected value of some loss function, L. We consider a situation where the loss function is not fully observed such that the evaluation of $\mathscr{X}$ must be based on an approximation of L. We denote the proxy for the true loss function by $\tilde{L}$ and seek conditions that will ensure that,

$$E(L(\mathscr{X}))\geqslant E(L(\mathscr{Y})) \quad \text{if and only if} \quad E(\tilde{L}(\mathscr{X}))\geqslant E(\tilde{L}(\mathscr{Y})) \quad \text{for all } \mathscr{X} \text{ and } \mathscr{Y}.$$

We formalize this idea in a setting where $\mathscr{X}$ and $\mathscr{Y}$ represent sequences of random variables that are being evaluated and compared in terms of their expected loss. In the context of volatility models the random variables, $\mathscr{X}$ and $\mathscr{Y}$, will denote two sequences of volatility forecasts that were produced by two competing models, whereas L and $\tilde{L}$ will be loss functions that are based on $\{\sigma_t^2\}_{t=1}^n$ and $\{\tilde{\sigma}_t^2\}_{t=1}^n$, respectively.

Definition 1 (*Set of alternatives*). The set of alternatives, $\mathbb{A}$, is a set of random sequences. A typical element of $\mathbb{A}$ is $\mathscr{X}(\omega)=\{X_1(\omega),X_2(\omega),\ldots\}$, which is defined on a probability space $(\Omega,\mathscr{F},P)$ and takes values in $(\mathbb{R}^\infty,\mathscr{B}_\infty)$, where $\mathscr{B}_\infty$ is the Borel σ-algebra on $\mathbb{R}^\infty$.[4]

[4] Thus $\mathscr{X}$ is a measurable mapping from $(\Omega,\mathscr{F})$ to $(\mathbb{R}^\infty,\mathscr{B}_\infty)$ and $\mathscr{B}_\infty$ is the smallest σ-algebra that contains all open subsets of $\mathbb{R}^\infty$ under the euclidian norm.

In the following we will often suppress the dependence on ω and simply write $\mathscr{X}$ in place of $\mathscr{X}(\omega)$. A statement that is said to hold *almost surely*, (a.s.), refers to the existence of a set $F \in \mathscr{F}$, with $P(F) = 1$, for which the statement is true for all $\omega \in F$.

Initially, we make the following assumptions about the two loss functions:

Assumption 1. Let L_t and $\tilde{L}_t$ be real functions, $t = 1, 2, \ldots$, and define the random variable $\hat{\psi}_n(\mathscr{X}) \equiv n^{-1}\sum_{t=1}^{n} \tilde{L}_t(X_t)$. For all $\mathscr{X} \in \mathbb{A}$ it holds that:

(i) $L_t(X_t)$ and $\tilde{L}_t(X_t)$ are integrable for all t;
(ii) $\psi(\mathscr{X}) \equiv \lim_{n\to\infty} n^{-1}\sum_{t=1}^{n} E[L_t(X_t)]$ and $\tilde{\psi}(\mathscr{X}) \equiv \lim_{n\to\infty} n^{-1}\sum_{t=1}^{n} E[\tilde{L}_t(X_t)]$ exist and are finite;
(iii) the limit $\hat{\psi}(\mathscr{X}) \equiv \lim_{n\to\infty} \hat{\psi}_n(\mathscr{X})$ exists and is finite a.s.

Thus ψ, $\tilde{\psi}$, and $\hat{\psi}_n$ serve as criteria for comparing alternatives in $\mathbb{A}$, such that ψ, $\tilde{\psi}$, and $\hat{\psi}_n$ induce preorderings on $\mathbb{A}$ that we denote by $\succcurlyeq$, $\overset{a}{\succcurlyeq}$, and $\overset{e}{\succcurlyeq}$, respectively. We shall refer to these preorderings as the *true preordering* the *approximate preordering*, and the *empirical preordering*, respectively. Note that $\overset{a}{\succcurlyeq}$ can be thought of as an approximation of $\succcurlyeq$, and that these two preorderings are non-stochastic preorderings, whereas $\overset{e}{\succcurlyeq}_n$, $n = 1, 2, \ldots$ is a sequence of stochastic preorderings on $\mathbb{A}$.

Definition 2 (*Preordering of alternatives*). For $\mathscr{X}, \mathscr{Y} \in \mathbb{A}$ we write: $\mathscr{X} \succcurlyeq \mathscr{Y}$ if $\psi(\mathscr{X}) \leqslant \psi(\mathscr{Y})$; $\mathscr{X} \overset{a}{\succcurlyeq} \mathscr{Y}$ if $\tilde{\psi}(\mathscr{X}) \leqslant \tilde{\psi}(\mathscr{Y})$; and we write $\mathscr{X} \overset{e}{\succcurlyeq}_n \mathscr{Y}$ if $\hat{\psi}_n(\mathscr{X}) \leqslant \hat{\psi}_n(\mathscr{Y})$, where $\overset{e}{\succcurlyeq}_n$ $n = 1, 2, \ldots$ is a sequence of stochastic preorderings.

It is easy to verify that the preorderings of Definition 2 are complete preorderings and we shall follow standard notation and write: '$\mathscr{X} \sim \mathscr{Y}$' if '$\mathscr{X} \succcurlyeq \mathscr{Y}$ and $\mathscr{X} \preccurlyeq \mathscr{Y}$' and '$\mathscr{X} \succ \mathscr{Y}$' if '$\mathscr{X} \succcurlyeq \mathscr{Y}$ and $\mathscr{X} \not\preccurlyeq \mathscr{Y}$' and similarly for the approximate preordering, $\overset{a}{\succcurlyeq}$, and the empirical preorderings, $\overset{e}{\succcurlyeq}_n$, $n = 1, 2, \ldots$.

The empirical preordering is obtained by averaging sample loss, and this is the only of the three preorderings that is directly observed by the econometrician. Naturally, the interesting question is whether the empirical preordering provide useful information about the true preordering. Next, we define concepts of equivalence between preorderings.

Definition 3 (*Equivalent and weakly equivalent*). Let $\succcurlyeq'$ and $\succcurlyeq''$ be preorderings and let $\succcurlyeq''_n$, $n = 1, 2, \ldots$ be a sequence of stochastic preorderings of $\mathbb{A}$. If it, for all $\mathscr{X}, \mathscr{Y} \in \mathbb{A}$, holds that

(a) $\mathscr{X} \succcurlyeq' \mathscr{Y} \Leftrightarrow \mathscr{X} \succcurlyeq'' \mathscr{Y}$, then $\succcurlyeq'$ and $\succcurlyeq''$ are equivalent on $\mathbb{A}$;
(b) $\mathscr{X} \succcurlyeq' \mathscr{Y} \Rightarrow P(\mathscr{X} \succcurlyeq''_n \mathscr{Y}) \underset{n\to\infty}{\to} 1$, then $\succcurlyeq''_n$ is asymptotically equivalent to $\succcurlyeq$ on $\mathbb{A}$;
(c) $\mathscr{X} \succ' \mathscr{Y} \Rightarrow P(\mathscr{X} \succ''_n \mathscr{Y}) \underset{n\to\infty}{\to} 1$, then $\succcurlyeq''_n$ is asymptotically weakly equivalent to $\succcurlyeq'$ on $\mathbb{A}$.

The difference between 'asymptotically equivalent' and 'asymptotically weakly equivalent' is that the former requires that "$\mathscr{X} \sim' \mathscr{Y}$ implies $\lim_{n\to\infty} P(\mathscr{X} \sim''_n \mathscr{Y}) = 1$", which is not guaranteed by the latter. The concept of asymptotic equivalence is useful for the analysis of the empirical preordering.

Remark 1. It should be noted that our definitions of equivalence are specific to the set of alternatives under considerations. Example, two preorderings that are equivalent on $\mathbb{A}$ need not be equivalent on a larger set of preorderings $\mathbb{A}'$.

Remark 1 has some implications for our analysis. Since $\mathbb{A}$ may consist of relatively few alternatives that are 'so' different that a slight difference between $\psi(\cdot)$ and $\tilde{\psi}(\cdot)$ need not affect the relative ranking of these alternatives. Below we derive sufficient conditions that ensure the equivalence of $\succcurlyeq$ and $\overset{\text{a}}{\succcurlyeq}$, and Remark 1 shows that these conditions need not be necessary conditions. We discuss this aspect in more details after Lemma 1 and will make reference to this observation in relation to some of our simulation results.

Lemma 1. *Define* $\gamma(\mathcal{X}) \equiv \psi(\mathcal{X}) - \tilde{\psi}(\mathcal{X})$ *and* $\gamma_n(\mathcal{X}) \equiv \tilde{\psi}(\mathcal{X}) - \hat{\psi}_n(\mathcal{X})$, (*where the latter is random*). (i) *If* $\delta(\mathcal{X}, \mathcal{Y}) \equiv \gamma(\mathcal{X}) - \gamma(\mathcal{Y}) = 0$ *for all* $\mathcal{X}, \mathcal{Y} \in \mathbb{A}$, *then* $\succcurlyeq$ *and* $\overset{\text{a}}{\succcurlyeq}$ *are equivalent*. (ii) *If* $\delta_n(\mathcal{X}, \mathcal{Y}) \equiv \gamma_n(\mathcal{X}) - \gamma_n(\mathcal{Y}) \overset{\text{a.s.}}{\to} 0$, *as* $n \to \infty$, *for all* $\mathcal{X}, \mathcal{Y} \in \mathbb{A}$, *then* $\overset{\text{e}}{\succcurlyeq}_n$ *is asymptotically weakly equivalent to* $\overset{\text{a}}{\succcurlyeq}$ *on* $\mathbb{A}$.

Lemma 1 shows that $\delta(\cdot, \cdot)$ can be interpreted as a measure of discrepancy between $\succcurlyeq$ and $\overset{\text{a}}{\succcurlyeq}$, so δ is directly related to the objective-bias. Similarly, $\delta_n(\cdot, \cdot)$ can be interpreted as a measure of discrepancy between $\overset{\text{a}}{\succcurlyeq}$ and $\overset{\text{e}}{\succcurlyeq}_n$ and is tied to the sampling error. As noted in Remark 1, a small deviation of δ from zero need not distort the equivalence of the preorderings, provided that the discrepancy is smaller than the difference in the expected average loss, i.e. $|\delta(\mathcal{X}, \mathcal{Y})| < |\psi(\mathcal{X}) - \psi(\mathcal{Y})|$ for all $\mathcal{X}, \mathcal{Y} \in \mathbb{A}$. So as we have commented on earlier, the conditions of Lemma 1 are sufficient conditions, but need not be necessary conditions.

2.1. Equivalence under parametric specification

In this subsection we adapt the theoretical framework to the comparison of volatility models by making additional assumptions about the loss function. Specifically we assume that L_t and $\tilde{L}_t$ have the same parametric form for all $t = 1, \ldots, n$ and only differ in terms of their 'parameters'.

Assumption 2. Let θ_t and $\tilde{\theta}_t$ denote two (possibly random) variables.

(i) For all $\mathcal{X} \in \mathbb{A}$, it holds that $L_t(X_t) \overset{\text{a.s.}}{=} L(\theta_t, X_t)$ and $\tilde{L}_t(X_t) \overset{\text{a.s.}}{=} L(\tilde{\theta}_t, X_t)$, where $\tilde{\theta}_t$ is a proxy for θ_t, $t = 1, 2, \ldots$.
Define $\eta_t \equiv \tilde{\theta}_t - \theta_t$ and let $\{\mathcal{F}_t\}$ be a filtration, such that for all $\mathcal{X} \in \mathbb{A}$, it holds that X_t and θ_t are $\mathcal{F}_{t-1}$-measurable, $t = 1, 2, \ldots$.

(ii) Either,
  (a) $L'(\theta, X) \equiv \partial L(\theta, X)/\partial\theta$ exists and does not depend on X; or
  (b) $L''(\theta, X) \equiv \partial^2 L(\theta, X)/\partial\theta\partial\theta'$ exists, does not depend on X, and $\{\eta_t, \mathcal{F}_t\}$ is a martingale difference sequence.

Assumption 2(i) requires that L_t and $\tilde{L}_t$ have the same parametric form such that the uncertainty about L_t is entirely expressed in terms of uncertainty about the parameter θ_t. We call θ_t and $\tilde{\theta}_t$ parameters although both may be random variables, as is the case in our application where $\theta_t = \sigma_t^2$ is the conditional variance. Assumptions 2(ii.a) and 2(ii.b) are assumptions about the functional form of L, which correspond to a linear and a quadratic form, respectively. The linearity, (ii.a), is unlikely to be satisfied in practical application, whereas the quadratic requirement is satisfied in some cases. An interesting observation is that several loss functions do not satisfy these requirements, so these loss functions may suffer from the objective-bias problem.[5]

Theorem 2. *Under Assumptions* 1(i–ii) *and* 2 *the true and the approximate preorderings,* $\succcurlyeq$ *and* $\overset{a}{\succcurlyeq}$, *are equivalent. Assumptions* 1(i–ii) *and* 2(i) *alone are not sufficient conditions for* $\succcurlyeq$ *and* $\overset{a}{\succcurlyeq}$ *to be equivalent.*

It is interesting to elaborate on the situation where Assumption 2(ii) does not hold. In this case we find that the larger is $\text{var}(\eta_t)$ the larger will the discrepancy, δ, tend to be, which makes it more likely that $\succcurlyeq$ and $\overset{a}{\succcurlyeq}$ disagree about the ranking of alternatives.

Corollary 3. *Let Assumptions* 1(i–ii) *and* 2(i) *hold, and suppose that Assumption* 2(ii) *is violated. Let the approximate preordering,* $\tilde{\psi}_\lambda$, *be defined by* $\tilde{\theta}_{\lambda,t} \equiv \theta_t + \lambda\eta_t$, $t = 1, \ldots, n$, *for some* $\lambda \in \mathbb{R}$, *where* $E(\eta_t|\mathcal{F}_{t-1}) \overset{\text{a.s.}}{=} 0$ *and* $\text{var}(\eta_t|\mathcal{F}_{t-1}) > 0$, a.s. *and define the discrepancy measure*

$$\delta_\lambda(\mathcal{X}, \mathcal{Y}) \equiv [\psi(\mathcal{X}) - \tilde{\psi}_\lambda(\mathcal{X})] - [\psi(\mathcal{Y}) - \tilde{\psi}_\lambda(\mathcal{Y})].$$

(i) *If the second derivative,* $\partial^2 L(\theta, X)/\partial\theta\partial\theta'$, *is bounded a away from zero, uniformly in* X, a.s., *then for some alternatives,* $\mathcal{X}$ *and* $\mathcal{Y}$, *it holds that* $|\delta_\lambda(\mathcal{X}, \mathcal{Y})| \to \infty$ *as* $\lambda \to \infty$.

(ii) *Under appropriate regularity conditions* (*see Assumption* A.1 *in the appendix*) *it holds that* $|\delta_\lambda(\mathcal{X}, \mathcal{Y})|$ *is strictly increasing in* $|\lambda|$ *for some* $\mathcal{X}, \mathcal{Y} \in \mathbb{A}$.

The broad message of Corollary 3 is that an increase in the conditional variance, $\text{var}(\theta_t - \tilde{\theta}_{\lambda,t}|\mathcal{F}_{t-1})$, is likely to cause an inconsistency between $\succcurlyeq$ and $\overset{a}{\succcurlyeq}$ when Assumption 2(ii) does not hold. See (A.3) of Lemma A.1 in the appendix for a detailed expression for $\delta_\lambda(\mathcal{X}, \mathcal{Y})$.

2.2. Consistency of the empirical preordering

Without knowledge about the probability measure, P, it is not possible to evaluate expected values, such as $E[L(\theta, X)] = \int L(\theta, X)\,\text{dP}$. So in practice it is not possible to rank alternatives in terms of $\succcurlyeq$ or $\overset{a}{\succcurlyeq}$. However, under suitable regularity conditions

[5] For example, an asymmetric loss function will typically fail to satisfy Assumption (ii.b). In practice, the relevant loss function will often be an asymmetric loss function, see Granger (1969) and Christoffersen and Diebold (2002) who characterize optimal forecasts under asymmetric loss.

the expected value can be approximated by a sample average, such that $\overset{e}{\succcurlyeq}_n$ resembles $\overset{a}{\succcurlyeq}$, and the equivalence of the two is formulated below.

Theorem 4. *Let Assumptions* 1 *and* 2(i) *hold and suppose that* $L(\tilde{\theta}_t, X_t),\ t = 1, 2, \ldots,$ *is stationary and ergodic for all* $\mathcal{X} \in \mathbb{A}$. *Then* $\overset{e}{\succcurlyeq}_n$ *is asymptotically weakly equivalent to* $\overset{a}{\succcurlyeq}$ *almost surely.*

Theorem 4 formulates one set of conditions that ensure the asymptotic (weak) equivalence of $\overset{e}{\succcurlyeq}_n$ and $\overset{a}{\succcurlyeq}$. The important aspect for obtaining this equivalence is that $n^{-1}\sum_{t=1}^{n} L(\tilde{\theta}_t, X_t) \xrightarrow{p} \tilde{\psi}(\mathcal{X}) = \lim_{n\to\infty} n^{-1}\sum_{t=1}^{n} E[L(\tilde{\theta}_t, X_t)]$, which can be obtained with different sets of assumptions. From a practical viewpoint the interesting question is whether $\overset{e}{\succcurlyeq}_n$ is asymptotically equivalent to the true preferences $\succcurlyeq$. So it is important that $\overset{a}{\succcurlyeq}$ is equivalent to $\succcurlyeq$, such that the empirical preordering can reveal the true ranking of alternatives, as defined by $\succcurlyeq$.

3. Comparison of volatility models

In this section, we show that our theoretical framework that concerns preorderings of stochastic sequences, yields valuable insight to the problem of comparing volatility models. We show that some, but not all, of the popular criteria for evaluating volatility models do satisfy the conditions we formulated in the previous section. For these criteria, it holds that an empirical ranking of alternatives is consistent for the intended ranking of alternatives. More importantly, we show that certain other criteria do not satisfy the needed conditions, so these may select an inferior model as the best. The result of Corollary 3 shows that the more precise is the proxy the less likely is the objective-bias to appear. In the context of the evaluation of volatility models this provides a strong argument for using the realized variance rather than a squared return as a proxy of σ_t^2.

The literature contains a vast number of studies that evaluate and compare volatility models, see, e.g., Poon and Granger (2003) that contains a review of 93 papers. Most papers apply a loss function, where model-based predictions, $\{h_t\}$, of the conditional variance, $\{\sigma_t^2\}$, is compared to proxies for the conditional variance, $\{\tilde{\sigma}_t^2\}$, typically squared daily returns. Common loss functions are: mean square (prediction) error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and logarithmic versions of these, where the loss functions take log-volatilities as the arguments. Another criterion for evaluating volatility models is the regression-based R^2-criterion that was suggested by Mincer and Zarnowitz (1969). This approach is frequently used in out-of-sample evaluation of forecasting models, where a proxy for the conditional variance is regressed on h_t and a constant. Pagan and Schwert (1990) noted that this regression is sensitive to 'outliers' of the proxies, assuming that estimation is made using the least squares method. This point was also made by Engle and Patton (2001). The problem is that the least squares estimates are strongly influenced by the largest realizations (outliers) of the proxy, $\tilde{\sigma}_t^2$. For this reason, Pagan and Schwert (1990) suggest to use a log-regression, where $\log(\tilde{\sigma}_t^2)$ is regressed on $\log(h_t)$ and a constant, because this regression is less sensitive to ‘outliers’.[6]

3.1. The framework

From a continuously compounded price process, $\{p_t\}$, $t \geqslant 0$, we define daily returns $r_t \equiv p_t - p_{t-1}$, $t = 1, 2, \ldots$ and the σ-algebra, $\mathscr{F}_t \equiv \sigma(r_t, r_{t-1}, \ldots)$, such that r_t is adapted to $\mathscr{F}_t$. We assume that $E(r_t^2) < \infty$ such that it is meaningful to define $\sigma_t^2 \equiv \mathrm{var}(r_t|\mathscr{F}_{t-1})$. For simplicity, we also assume that $E(r_t|\mathscr{F}_{t-1}) = 0$, such that $\{r_t, \mathscr{F}_t\}$ is a martingale difference sequence.

We consider volatility models that describe the variation in the conditional variance σ_t^2, $t = 1, 2, \ldots$. Model i yields a sequence, $h_{i,t}$, $t = 1, \ldots, n$, where $h_{i,t}$ is a ‘predictor’ of σ_t^2. So our set of alternatives, $\mathbb{A}$, consists of the sequences, $\mathscr{H}_i = (h_{i,1}, \ldots, h_{i,n})$, where the subscript $i = 1, \ldots, m$ indexes the different volatility models.

An immediate obstacle for the evaluation of volatility models is the fact that σ_t^2 is unobserved. A solution is to substitute a proxy for σ_t^2, such as squared daily returns, $\tilde{\sigma}_t^2 = r_t^2$, or $\tilde{\sigma}_t^2 = (r_t - \hat{\mu}_t)^2$, where $\hat{\mu}_t$ is an estimate of the conditional mean, $E(r_t|\mathscr{F}_{t-1})$. It is not surprising that the squared daily return is a rather noisy measure of σ_t^2. In fact, when volatility models are evaluated using squared daily returns, it results in (what appears to be) a very poor performance, see Andersen and Bollerslev (1998). A better choice for $\tilde{\sigma}_t^2$ is a measure that incorporates the additional information that intraday data have to offer about σ_t^2. Proxies of this kind include the range-based estimators of σ_t^2, which are based on the “open”, “low”, “high”, and “close” prices for a given trading day, see e.g. Alizadeh et al. (2002) and references therein. A better, but also more computational intensive, measure of daily volatility is the RV, see e.g., Andersen and Bollerslev (1998) and Andersen et al. (2001, 2003). This measure is also known as the realized volatility; in this paper we follow Barndorff-Nielsen and Shephard (2002a,b) and refer to RV as the ‘realized variance’. The RV is constructed by taking the sum of squared intraday returns that yields an unbiased measure of the conditional variance, σ_t^2, under suitable regularity conditions.

It is now clear that this problem fits into the framework of the previous section, where σ_t^2 and $\tilde{\sigma}_t^2$ play the role of θ_t and $\tilde{\theta}_t$, respectively, and where the set of alternative, $\mathbb{A}$, is given by the sequences of forecasts, $\mathscr{H}_i = (h_{i,1}, \ldots, h_{i,n})$, $i = 1, \ldots, m$.

3.2. Evaluation based on loss functions

We consider two loss functions, the MSE loss function, $L(\sigma_t^2, h_t) = (\sigma_t^2 - h_t)^2$, and the logarithmic version of it (MSE*), which is given by $L(\sigma_t^2, h_t) = [\log(\sigma_t^2) - \log(h_t)]^2$. Our motivation for analyzing the properties of these loss functions is that

[6] An alternative way to deal with this problem is to use a ‘robust’ estimation method instead of the least squares method, as was pointed out by an anonymous referee.

these have been used in the literature. So we work out the specific details in relation to the theoretical framework of the previous section.

3.2.1. MSE: mean squared error loss

First consider the MSE loss function $L(\sigma_t^2, h_t) = (\sigma_t^2 - h_t)^2$. A Taylor expansion of the approximating loss function, $L(\tilde{\sigma}_t^2, h_t) = (\tilde{\sigma}_t^2 - h_t)^2$, about σ_t^2 is given by

$$L(\tilde{\sigma}_t^2, h_t) = L(\sigma_t^2, h_t) + 2(\sigma_t^2 - h_t)\eta_t + \eta_t^2,$$

where $\eta_t = \tilde{\sigma}_t^2 - \sigma_t^2$. So Assumption 2(ii.b) is satisfied whenever $\tilde{\sigma}_t^2$ is conditionally unbiased for σ_t^2, and we can conclude that $L(\sigma^2, \cdot)$ and $L(\tilde{\sigma}^2, \cdot)$ induce the same preordering in this case. In particular we have that

$$\arg\min_{\mathbb{A}} E\left[n^{-1}\sum_{t=1}^{n} L(\sigma_t^2, h_{i,t})\right] = \arg\min_{\mathbb{A}} E\left[n^{-1}\sum_{t=1}^{n} L(\tilde{\sigma}_t^2, h_{i,t})\right],$$

such that the volatility model with the smallest population loss (the best model) coincides with the model that minimizes the expected approximate loss.

3.2.2. MSE*: mean squared log relative error loss

Next, consider the MSE* loss function, $L(\sigma_t^2, h_t) = [\log(h_t/\sigma_t^2)]^2 = [\log(\sigma_t^2) - \log(h_t)]^2$. The first two derivatives that are relevant for our analysis are given by

$$\frac{\partial L}{\partial \sigma_t^2} = 2\,\frac{\log(\sigma_t^2/h_t)}{\sigma_t^2} \quad \text{and} \quad \frac{\partial^2 L}{\partial \sigma_t^2 \partial \sigma_t^2} = 2\,\frac{1 - \sigma_t^2 \log(\sigma_t^2/h_t)}{\sigma_t^4},$$

which do not satisfy Assumption 2(ii).

Given the failure of this loss function (MSE*) to satisfy Assumption 2(ii), we conclude that an evaluation that is based on this loss function may result in an objective-bias, such that the ‘best’ in terms of the $E[L(\tilde{\sigma}_t^2, \cdot)]$ need not be identical to the ‘best’ as defined by the true criterion, $E[L(\sigma_t^2, \cdot)]$. So in order to avoid an objective-bias using this loss function it is not sufficient that $\tilde{\sigma}_t^2$ is conditionally unbiased for σ_t^2, is employed ($E(\eta_t|\mathscr{F}_{t-1}) = 0$). An objective-bias is more likely to arise the larger is $E(\eta_t^2|\mathscr{F}_{t-1})$, in fact a large conditional variance causes the approximate evaluation to favor models for which $E[\log h_t]$ is relatively small. This can be seen from the Taylor expansion of this loss function. Naturally, if $\log(\tilde{\sigma}_t^2)$ is conditionally unbiased for $\log(\sigma_t^2)$, then we have a situation that is identical to that of the MSE loss function, where Assumption 2(ii) is satisfied.

3.3. Regression based evaluation

A popular alternative to using loss functions for evaluating volatility models, is to use the R^2 of the Mincer–Zarnowitz regressions that have the form

$$\varphi(\tilde{\sigma}_t^2) = \alpha + \beta\varphi(h_t) + u_t, \quad t = 1, \ldots, n. \tag{1}$$

Common choices for the φ-function include the identity, $\varphi(x) = x$, and the logarithmic transformation, $\varphi(x) = \log(x)$. These result in the two regression equations

$$\tilde{\sigma}_t^2 = \alpha + \beta h_t + u_t, \quad t = 1, \ldots, n, \tag{2}$$

$$\log \tilde{\sigma}_t^2 = \alpha + \beta \log h_t + u_t, \quad t = 1, \ldots, n, \tag{3}$$

respectively. As mentioned earlier, Pagan and Schwert (1990) have argued in favor of the log-regression (3) because the level-regression (2) may be sensitive to extreme (large) values of r_t^2.

The R^2-criterion induces the same preordering of alternatives as[7]

$$\psi(\mathscr{H}) \equiv \text{cov}(\varphi(\sigma_t^2), \varphi(h_t))[\text{var}(\varphi(h_t))]^{-1}\text{cov}(\varphi(h_t), \varphi(\sigma_t^2)), \tag{4}$$

where $\mathscr{H}$ is generic for a sequence $(h_{i,1}, h_{i,2}, \ldots)$, $i = 1, \ldots, m$. Similarly, the substitution of $\tilde{\sigma}_t^2$ for σ_t^2 in the regression equation results in an R^2-criterion that is equivalent to

$$\tilde{\psi}(\mathscr{H}) \equiv \text{cov}(\varphi(\tilde{\sigma}_t^2), \varphi(h_t))[\text{var}(\varphi(h_t))]^{-1}\,\text{cov}(\varphi(h_t), \varphi(\tilde{\sigma}_t^2)). \tag{5}$$

The underlying criterion of (4) and (5) does not fit directly into our framework of Section 2. Nevertheless, we shall establish conditions that ensure that $\psi(\mathscr{H})$ and $\tilde{\psi}(\mathscr{H})$ induce the same preorderings. The following assumption contains the relevant conditions that are akin to those of the previous section.

Assumption 3. Let $\eta_t \equiv \sigma_t^2 - \tilde{\sigma}_t^2$. The φ-function in (1) is infinite differentiable[8] and it holds that

$$E((\eta_t)^j | \mathscr{F}_{t-1}) \frac{\partial^j \varphi(x)}{\partial (x)^j}\bigg|_{x=\sigma_t^2} \overset{\text{a.s.}}{=} c_j$$

for some constant $c_j, \in \mathbb{R}$, for all $t = 1, 2, \ldots,$ and all $j = 1, 2, \ldots$.

This assumption is simple to interpret in the special cases that are seen in applied work, such as the cases where φ is linear or logarithmic. Below, we derive the specific requirements of Assumption 3 for the three cases where φ is linear, quadratic, and logarithmic.

Theorem 5. *The criteria, $\psi(\mathscr{H})$ and $\tilde{\psi}(\mathscr{H})$, are equivalent under Assumption* 3.

Linear φ. Suppose that φ is linear, such that the first derivative, φ', is constant and higher order derivatives are zero. Assumption 3 simply requires that $E(\eta_t | \mathscr{F}_{t-1})$ is constant (almost surely).

So the R^2-criterion of (2) is not affected by a conditional bias of $\tilde{\sigma}_t^2$. This robustness is obtained at a cost, because the criterion is unable to distinguish between the volatility models, $\mathscr{H}_1 = \{h_t\}$ and $\mathscr{H}_2 = \{a + bh_t\}$, for arbitrary values of a and $b \neq 0$.

[7] The R^2-criterion is given by $\psi(\mathscr{H})/\text{var}(\varphi(\sigma_t^2))$ which induces the same preordering (ranking) as $\psi(\mathscr{H})$, because $\text{var}(\varphi(\sigma_t^2))$ is constant across alternatives.

[8] By infinite differentiable is meant that the pth derivative of φ exists for any integer, p.

Quadratic φ. Suppose that φ is quadratic, such that the second order derivative, φ'', is constant and higher-order derivatives are all zero. The conditions in Assumption 3 simplify to $E(\eta_t|\mathscr{F}_{t-1})\varphi'(\sigma_t^2)$ and $E(\eta_t^2|\mathscr{F}_{t-1})$ being equal to some constants (almost surely).

Logarithmic φ. Suppose that the models are compared using the R^2s from regressions in the form of (3). If the proxy, $\tilde{\sigma}_t^2$, is conditionally unbiased for σ_t^2, then it is unlikely that the R^2 from the feasible regressions, (3), will induce the same ranking of volatility models as the R^2 from the infeasible regressions $\log \sigma_t^2 = a + b \log h_t + u_t$, because Assumption 3 is not satisfied. This is clearly an unfortunate property of the log-regression criterion, in particular if squared returns, r_t^2, are substituted for the unobserved conditional variance, σ_t^2, as this leads to an $\eta_t = r_t^2 - \sigma_t^2$ with a large variance. Analogous to the case with an additive loss function, see Corollary 3, the distortion becomes more severe the larger is the conditional variance of η_t. The last point is easily seen from an inspection of the proof of Theorem 5.

Suppose now that $\tilde{\sigma}_t^2 = (1 - \upsilon_t)\sigma_t^2$, for some random variable, υ_t,[9] where the conditional moments of υ_t, given by $\kappa_j \equiv E(\upsilon_t^j|\mathscr{F}_{t-1})$, $j = 1, 2, \ldots,$ are finite and constant. This implies that the conditional bias of $\log \tilde{\sigma}_t^2$, relative to $\log \sigma_t^2$, is constant. The measurement error is then given by $\eta_t = \upsilon_t \sigma_t^2$, and $E(\eta_t^j|\mathscr{F}_{t-1}) = E(\upsilon_t^j|\mathscr{F}_{t-1})(\sigma_t^2)^j$. Since $[\partial^j \log(x)]/\partial x^j = [(-1)^{j-1}/(j-1)!]x^{-j}$, we have that

$$E((\eta_t)^j|\mathscr{F}_{t-1})\frac{\partial^j \varphi(x)}{\partial (x)^j}\bigg|_{x=\sigma_t^2} = \kappa_j \frac{(-1)^{j-1}}{(j-1)!},$$

which is constant for all $j = 1, 2, \ldots$. So a measurement error of this kind will not create an objective-bias for this criterion. This observation has practical relevance due to the results by Barndorff-Nielsen and Shephard (2002a,b) who have shown that the variance of the RV as an estimator of σ_t^2, increases with σ_t^2. See also Meddahi (2002).[10]

4. Empirical and simulation-based comparisons of ARCH-type models

In this section we explore the empirical relevance of the theoretical results that we derived in the previous two sections. This is done by evaluating and comparing ARCH-type volatility models using real data (IBM stock returns) and a large number of artificial data sets.

[9] One may impose that $v_t \leqslant 1$ (a.s.) to ensure that $\tilde{\sigma}_t^2$ is non-negative (a.s.).

[10] More precisely: the variance of the realized variance increases with the integrated variance, where the latter is conditionally unbiased for the conditional variance, σ_t^2.

4.1. Empirical comparison based on IBM equity returns

We evaluate and compare eight ARCH-type volatility models using IBM stock price data that were extracted from the Trade and Quote (TAQ) database.[11] The realized variance that is used in our evaluation was calculated from intraday returns of mid-quotes between 9:30 AM and 4 PM. The sample period is January 3, 1995 through February 21, 2002, which adds up to a total of 1795 trading days. The models were estimated using the first 1250 observations (up until December 13, 1999) and the remaining 545 observations were used for the out-of-sample evaluation and comparison.

Our set of competing ARCH-type models comprises the ARCH model by Engle (1982), the GARCH model by Bollerslev (1986), the threshold GARCH model (Thr.-GARCH) by Zakoian (1994), the EGARCH model by Nelson (1991), the APARCH model that was proposed by Ding et al. (1993), the FIGARCH model suggested by Baillie et al. (1996), and the FIAPARCH model by Tse (1998).

Each model is estimated using daily close-to-close returns. In the evaluation we apply the MSE and the MSE* loss functions and two Mincer–Zarnowitz regressions, the level-regression (2) and the log-regression (3), and we employ three different proxies for σ_t^2 that differ in terms of the variance of their associated measurement errors. Two of the proxies are based on the realized variance, RV_t, that is calculated from high-frequency data (mid-quotes) during the hours that the market is open for trading. Our measure of RV_t was calculated with the Fourier method by Malliavin and Mancino (2002). See Appendix A for details.

Our three proxies for σ_t^2 are given by $\tilde{\sigma}^2_{[\mathrm{sc.RV}]t} \equiv \hat{c}\,\mathrm{RV}_t$, $\tilde{\sigma}^2_{[\mathrm{RV+on}]t} \equiv \mathrm{RV}_t + (p_t^{\mathrm{open}} - p_{t-1}^{\mathrm{close}})^2$, and $\tilde{\sigma}^2_{[\mathrm{sq.ret}]t} \equiv (p_t^{\mathrm{close}} - p_{t-1}^{\mathrm{close}})^2$. The constant, $\hat{c} \equiv n^{-1}\sum_{t=1}^{n} r_t^2/\mathrm{RV}_t$, is a scaling factor that accounts for the fact that RV_t is a measure of volatility for the open-to-close period. So RV_t is not a proper proxy of σ_t^2 because it does not span the full day (close-to-close). The second proxy, $\tilde{\sigma}^2_{[\mathrm{RV+on}]t}$, accounts for the volatility during the close-to-open time period, by adding the squared over-night return. Finally $\tilde{\sigma}^2_{[\mathrm{sq.ret}]t}$ is simply the squared (inter-day) return, r_t^2. See Hansen and Lunde (2004a) for additional details about these proxies. It is worth to note that

$$\mathrm{var}(\tilde{\sigma}^2_{[\mathrm{sc.RV}]t} - \sigma_t^2) \leqslant \mathrm{var}(\tilde{\sigma}^2_{[\mathrm{RV+on}]t} - \sigma_t^2) \leqslant \mathrm{var}(\tilde{\sigma}^2_{[\mathrm{sq.ret}]t} - \sigma_t^2),$$

under quite reasonable assumptions. So $\tilde{\sigma}^2_{[\mathrm{sq.ret}]t}$ is more likely to create an objective-bias for the criteria that do not satisfy the ‘equivalence conditions’, where ‘equivalence conditions’ refer to the conditions that ensure the equivalence of $\succcurlyeq$ and $\overset{\mathrm{a}}{\succcurlyeq}$.

[11] The TAQ database contains all trades and quotes in the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and National Association of Securities Dealers Automated Quotation (Nasdaq) securities.

Table 1
Mean squared errors loss for GARCH models of IBM stock returns

Model	MSE (level)			MSE (log)		
	$\mathrm{MSE}_{\mathrm{sc.RV}}$	$\mathrm{MSE}_{\mathrm{RV+on}}$	$\mathrm{MSE}_{\mathrm{sq.ret}}$	$\mathrm{MSE}^{\star}_{\mathrm{sc.RV}}$	$\mathrm{MSE}^{\star}_{\mathrm{RV+on}}$	$\mathrm{MSE}^{\star}_{\mathrm{sq.ret}}$
ARCH(1)	39.924	168.30	305.06	0.492	0.599	**11.882**
	(7.159)	(126.67)	(150.21)	(0.030)	(0.046)	(3.639)
GARCH(1,1)	30.722	159.77	297.01	0.298	0.479	12.441
	(5.820)	(120.33)	(142.98)	(0.019)	(0.033)	(3.718)
EGARCH(1,1)	25.434	155.25	289.37	**0.260**	**0.459**	12.506
	(5.389)	(120.01)	(142.64)	(0.016)	(0.032)	(3.733)
A-PARCH(1,1)	**23.358**	**153.66**	**286.52**	0.269	0.485	12.644
	(5.147)	(120.01)	(142.54)	(0.015)	(0.031)	(3.736)
THR-GARCH(1,1)	24.711	155.00	288.49	0.261	0.471	12.594
	(5.111)	(119.39)	(141.86)	(0.015)	(0.031)	(3.741)
FIGARCH(0,0)	31.460	161.89	299.03	0.338	0.543	12.573
	(6.376)	(123.15)	(146.01)	(0.020)	(0.036)	(3.724)
FIGARCH(1,1)	32.057	161.89	299.51	0.344	0.564	12.765
	(6.350)	(121.08)	(143.76)	(0.021)	(0.036)	(3.744)
FIAPARCH(1,1)	24.334	155.53	288.11	0.293	0.533	12.883
	(4.854)	(119.69)	(141.96)	(0.017)	(0.032)	(3.768)

This table reports the sample loss for eight volatility models, where the evaluation is based on three different proxies for the conditional variance σ_t^2. The proxies are denoted by $\tilde{\sigma}^2_{[\mathrm{sc.RV}]t}$, $\tilde{\sigma}^2_{[\mathrm{RV+on}]t}$, and $\tilde{\sigma}^2_{[\mathrm{sq.ret}]t}$, which refer to 'the scaled RV$_t$', 'the RV$_t$ plus the overnight return', and 'the squared close-to-close return', respectively. The MSE-criterion is based on the difference between $\tilde{\sigma}_t^2$ and the model-based prediction of volatility, h_t, whereas the MSE*-criterion is based on the difference between $\ln \tilde{\sigma}_t^2$ and $\ln h_t$. Bold font identifies the best sample performance in each column, and the standard errors of the MSEs are given in parentheses.

4.2. Empirical results

We shall refer to the MSE loss function and the regression in levels as *robust criteria*, because they satisfy the equivalence conditions. The two other criteria, MSE* and the $R^{\star 2}$ from the regression in logs, are referred to as *sensitive criteria*, because they do not satisfy the relevant conditions.

The empirical results of the evaluation of the eight volatility models are given in Tables 1 and 2. Table 1 contains the results based of the two loss functions and the analogous results for the regression-based criteria are presented in Table 2. The empirical results corroborate our theoretical results. It is striking that the robust criteria point to the same model as the best volatility model, for all three choices of $\tilde{\sigma}_t^2$, whereas the sensitive criteria point to different models. The MSE* points to the ARCH model as the best volatility models when a squared return is substituted for σ_t^2. Since it is well known that the ARCH model is unable to fully capture the persistence in the conditional variance, it is very unlikely that this model is the best model. Instead this result must be attributed to chance that only occurs because the squared return is a noisy measure of σ_t^2. Another aspect that stands out from Table 1 is the large difference in standard errors across the three proxies—even when these are measured relative to the corresponding sample losses. The gains from using the

Table 2
R^2 from Mincer–Zarnowitz Regressions for GARCH models of IBM stock returns

Model	Level regression			Log-regression		
	$R^2_{\text{sc.RV}}$	$R^2_{\text{RV+on}}$	$R^2_{\text{sq.ret}}$	$R^{\star 2}_{\text{sc.RV}}$	$R^{\star 2}_{\text{RV+on}}$	$R^{\star 2}_{\text{sq.ret}}$
ARCH(1)	0.067	0.010	0.003	0.131	0.102	0.009
GARCH(1,1)	19.907	4.328	1.808	43.301	35.048	1.531
EGARCH(1,1)	30.575	6.365	3.596	52.007	42.366	**2.625**
A-PARCH(1,1)	**36.052**	**7.329**	**4.525**	52.182	42.816	2.483
THR-GARCH(1,1)	32.425	6.665	3.871	**52.442**	42.700	2.622
FIGARCH(0,0)	16.867	3.112	1.191	37.827	30.773	1.614
FIGARCH(1,1)	16.338	3.420	1.226	38.245	31.471	1.314
FIAPARCH(1,1)	33.784	6.737	4.037	51.791	**42.931**	2.532

This table reports the R^2s from the Mincer–Zarnowitz regressions. In the level regressions we have regressed a proxy, $\tilde{\sigma}_t^2$, on a model forecast, h_t, and a constant, whereas $\ln \tilde{\sigma}_t^2$ was regressed on $\ln h_t$ and a constant in the log-regressions. Three different proxies for σ_t^2 were used: $\tilde{\sigma}^2_{[\text{sc.RV}]t}$ (scaled RV_t), $\tilde{\sigma}^2_{[\text{RV+on}]t}$ (RV_t plus the overnight return), and $\tilde{\sigma}^2_{[\text{sq.ret}]t}$ (squared close-to-close return, r_t^2). Bold font identifies the best sample performance in each column. Note that the robust R^2-criterion points to the same model as 'best' for all proxies, whereas the sensitive $R^{\star 2}$-criterion points to different models as 'best'.

most precise estimator (the scaled RV) are very large. So in addition to the benefit that an objective-bias might be avoided, the precise proxy leads to a more informative comparison with tighter confidence intervals for the relative performances. Although the standard errors of the performance measures for each of the models are relatively large, it is quite likely that the standard errors of relative performances (between pairs of models) are small, because $\text{var}(L_{i,t} - L_{j,t}) = \text{var}(L_{i,t}) + \text{var}(L_{j,t}) - 2\text{cov}(L_{i,t}, L_{j,t}) \approx 2\text{var}(L_{i,t})(1-\rho)$, where $\rho \equiv \text{corr}(L_{i,t}, L_{j,t})$ and we expect ρ to be large (close to one).

Next, we turn to a simulation study to compare the criteria in question. This has the advantage that the conditional variance is known to us and so is the true data generating process.

4.3. Simulation based comparison

We generate artificial data from a GARCH(1,1) model, an EGARCH(1,1) model, and an APARCH(1,1) model, where we used the parameter estimates from the analysis of IBM data as the values for the population parameters of these models.[12]

$$\text{GARCH}(1,1): \quad \sigma_t^2 = \underset{(0.007)}{0.043} + \underset{(0.009)}{0.121}\,\varepsilon_{t-1}^2 + \underset{(0.012)}{0.846}\,\sigma_{t-1}^2,$$

$$\text{EGARCH}(1,1): \quad \ln\sigma_t^2 = -\underset{(0.012)}{0.121} + \underset{(0.016)}{0.163}(|\varepsilon_{t-1}| - \underset{(0.071)}{0.478}\,\varepsilon_{t-1})\sigma_{t-1}^{-1} + \underset{(0.004)}{0.975}\,\ln\sigma_{t-1}^2,$$

$$\text{APARCH}(1,1): \quad \sigma_t^{\underset{(0.095)}{0.433}} = -\underset{(0.004)}{0.025} + \underset{(0.008)}{0.078}(|\varepsilon_{t-1}| - \underset{(0.066)}{0.555}\,\varepsilon_{t-1})^{\underset{(0.095)}{0.433}} + \underset{(0.007)}{0.918}\,\sigma_{t-1}^{\underset{(0.095)}{0.433}}.$$

[12] The standard errors of the estimated parameters are listed in brackets.

The results are presented in Figs. 1–3. Fig. 1 contains the results from the artificial data that where generated using the GARCH model, and Figs. 2 and 3 contain the analogous results from the data that were generated with the EGARCH model and the APARCH model, respectively. The simulations are based on 500 trials. Five thousand observations were used for the initial estimation and an additional 100, 250, or 500 observations were used for the recursive out-of-sample evaluation.

From Fig. 1 it can be seen that the true GARCH model often has the smallest loss when the true conditional variance is used in the evaluation. This holds for both loss functions. However, when we substitute the squared return for the conditional variance, the GARCH model rarely has the smallest sample loss and there is a noticeable difference between the two loss functions. When the evaluation is based on 250 and 500 observations the GARCH is more likely to have a smaller sample loss than any other model using the MSE loss function. This is in sharp contrast to the results for the MSE* loss function, where the Component Thr.-GARCH model very frequently out-performs all other models including the GARCH(1,1) model. The results from the simulations based on the EGARCH model that are presented in Fig. 2 similarly show a pronounced objective-bias when the EGARCH is the true data generating model. Again we find that the component Thr.-GARCH spuriously outperforms the true model when the evaluation is based on the noisy proxy for σ_t^2.

A slightly different picture emerges from Fig. 3, where the data are generated from the APARCH model. Here we find that the APARCH is indeed the model that most frequently is found to be the best, even when the noisy proxy is used. The reason is that the APARCH with the present configuration generates a dynamic in the conditional variance that is difficult for the other models to imitate. Note that the point estimate for the power-parameter is 0.433 which differs quite a bit from 2.000 which is tacitly imposed by most volatility models. So the true model is sufficiently different from the other models that the objective-bias does not become an issue for the best model. See Remark 1 and the discussion that followed Remark 1. While the objective bias may not be relevant for the best model it may be an issue for other (lower ranked) models. For example, the results based the true values of σ_t^2 indicate that EGARCH is the second best model. This is not too surprising because the EGARCH can accommodate a leverage effect that is similar to that of the APARCH model. While the EGARCH model is ranked second (in terms of the frequency at which it performs best) when the σ_t^2 is used in the evaluation, it is never ranked better than third when the MSE* criterion is employed in combination with the noisy proxy. This suggests that an objective-bias may affect this (lower ranked) model.

The important message from our empirical and simulation-base results is that the combination of 'a non-robust criterion' and 'a noisy proxy' can be devastating for model selection. In fact, our results indicate that applied users would tend to favor the wrong model in this case.

5. Summary

It is well-documented that the choice of proxy can affect the quantitative assessment of a given volatility model, see Andersen and Bollerslev (1998). In this

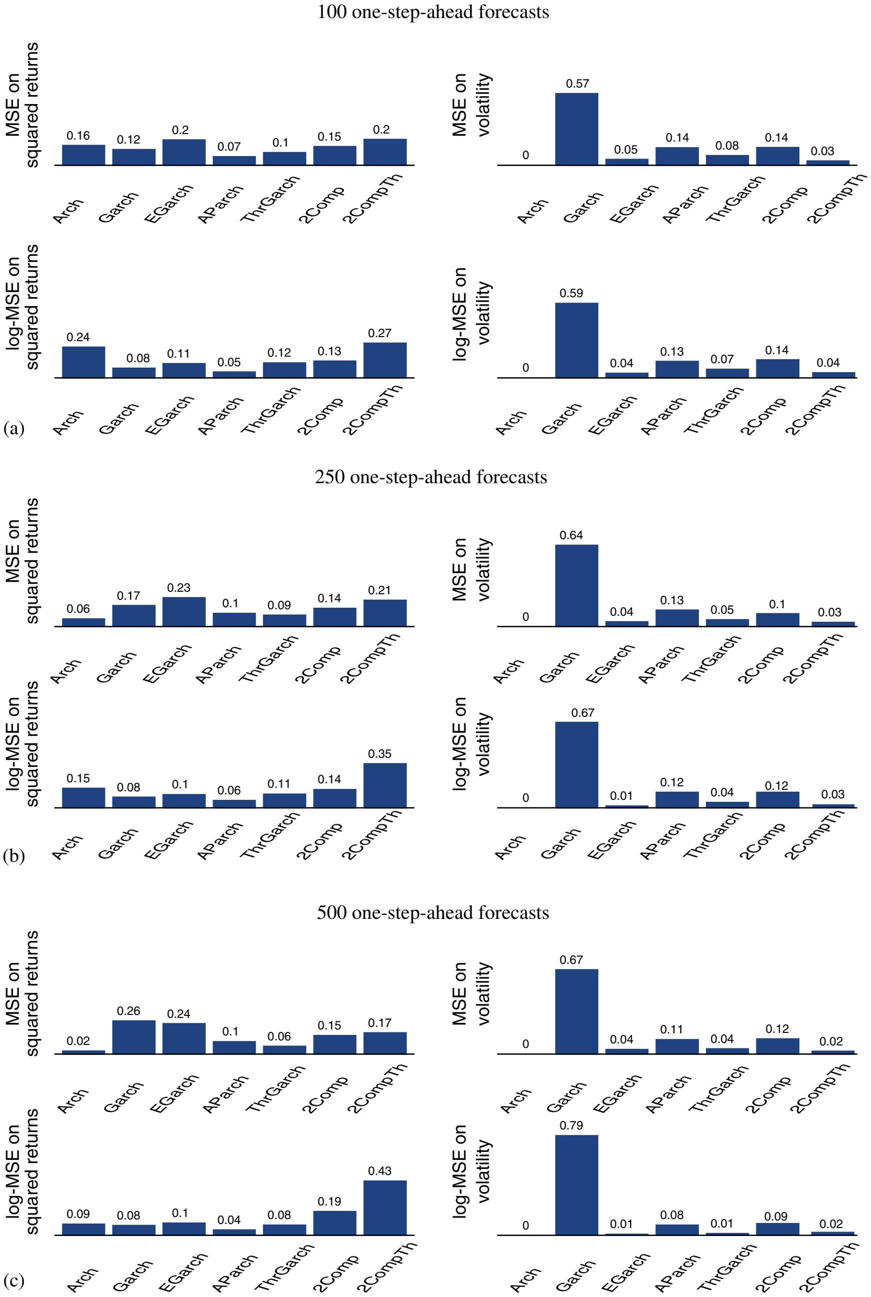
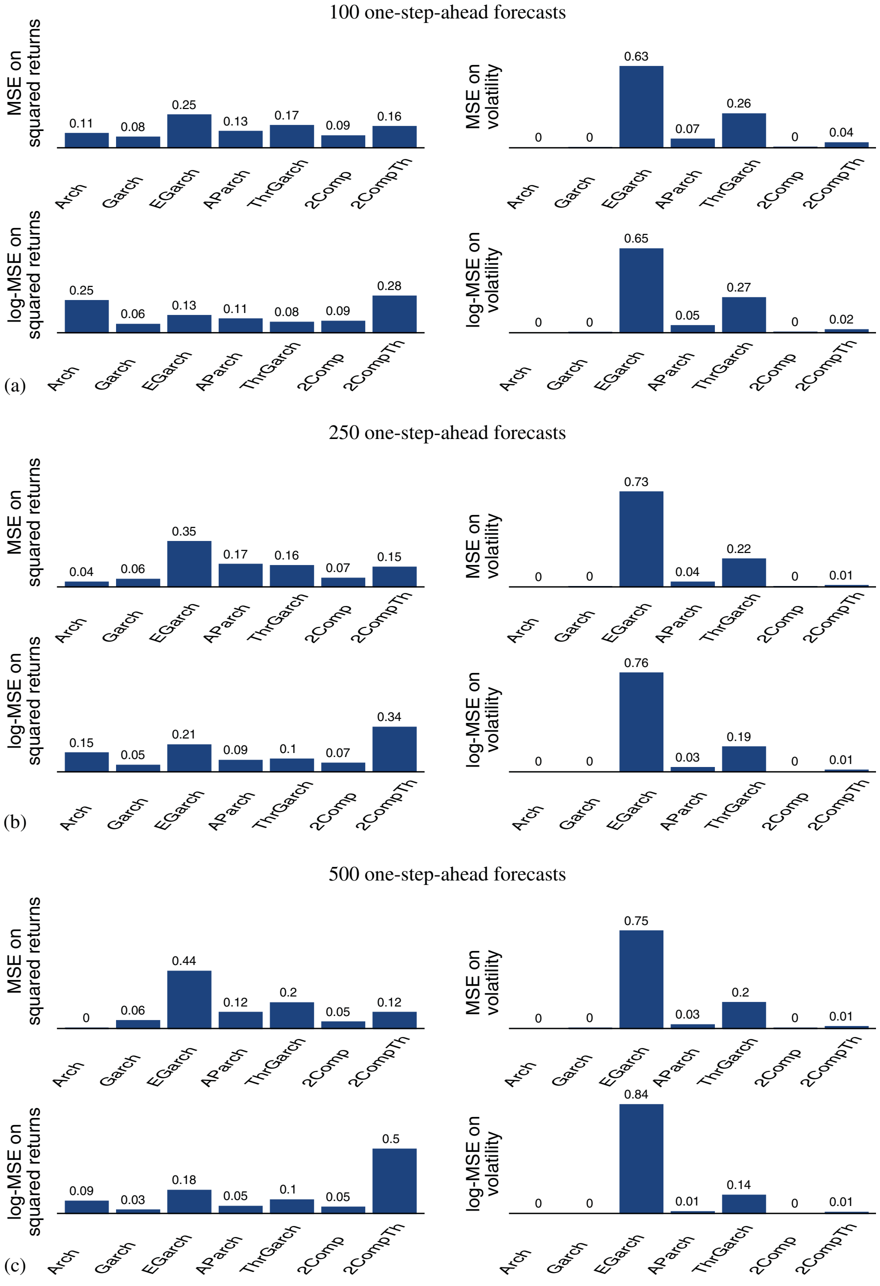


Fig. 1. The figure shows the frequency that each model had the smallest out-of-sample loss. The artificial data was simulated from a GARCH(1,1) model. The three panels contain the results for 100, 250, and 500 out-of-sample observations. The first (second) column contains the evaluation based on squared returns (the true conditional variance); and the MSE (MSE*) loss function was applied in rows 1, 3, and 5 (2, 4, and 6).


100 one-step-ahead forecasts

(a)

250 one-step-ahead forecasts

(b)

500 one-step-ahead forecasts

(c)

Fig. 2. The figure shows the frequency that each model had the smallest out-of-sample loss. The artificial data was simulated from an EGARCH(1,1) model. The three panels contain the results for 100, 250, and 500 out-of-sample observations. The first (second) column contains the evaluation based on squared returns (the true conditional variance); and the MSE (MSE*) loss function was applied in rows 1, 3, and 5 (2, 4, and 6).

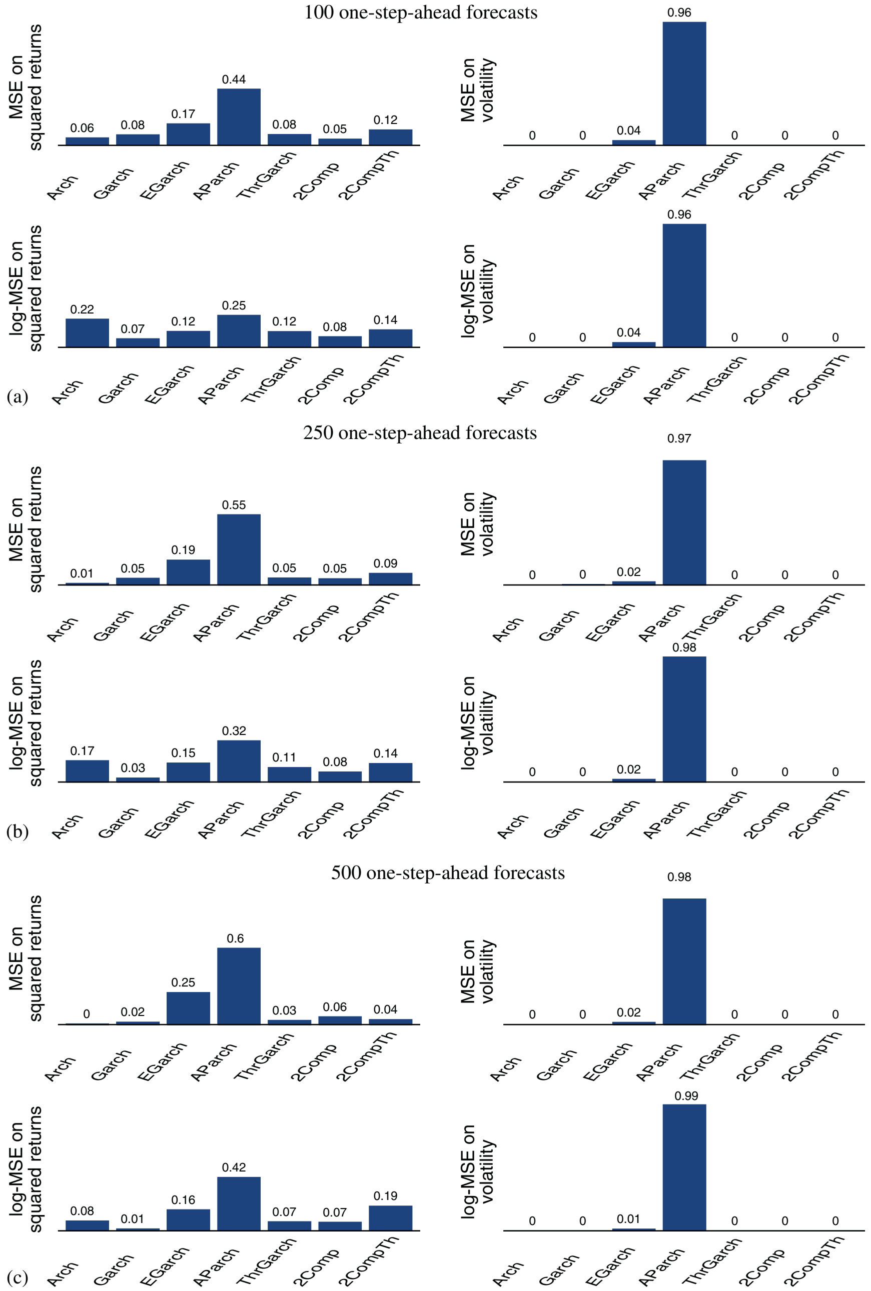


Fig. 3. The figure shows the frequency that each model had the smallest out-of-sample loss. The artificial data was simulated from an APARCH(1,1) model. The three panels contain the results for 100, 250, and 500 out-of-sample observations. The first (second) column contains the evaluation based on squared returns (the true conditional variance); and the MSE (MSE*) loss function was applied in rows 1, 3, and 5 (2, 4, and 6).

paper we have analyzed how the substitution may affect the qualitative assessment. Specifically we have shown that some criteria that have been used for the evaluation of volatility models can be distorted by the substitution of a proxy for the latent population measure of volatility. The substitution of a squared return for the conditional variance, σ_t^2, is our leading example.

We established our results in a general theoretical framework that involved three preorderings of the alternatives to be compared: the true preordering, an approximate preordering, and an empirical preordering. In this paper, we have studied the discrepancies between these preorderings and separated the sampling error from the objective-bias. While the sampling error makes it more difficult to tell good and bad models apart, the objective-bias is more severe for the empirical analysis, because it causes the empirical ranking to be inconsistent for the true ranking. We have derived conditions that ensure that the true and the approximate preorderings are equivalent, and conditions that ensure that the empirical preordering is asymptotically weakly equivalent to the approximate preordering. Thus, if both sets of conditions hold, the empirical preordering will asymptotically coincide with the true preordering, which is to be desired in practice.

In the context of evaluation and comparison of volatility models we have shown that commonly used criteria, such as the logarithmic version of the Mincer–Zarnowitz regression, do not meet the conditions required for a consistent ranking of volatility forecasting models. So these criteria may result in a ranking of models that is different from the intended ranking.

In an empirical analysis of IBM stock returns we have compared eight volatility models using three different proxies for the unobserved conditional variance. The empirical results revealed the practical relevance of the theoretical results. The 'sensitive' criteria that are subject to the objective-bias problem, pointed to different models as the best when different proxies were used. This was in sharp contrast to the 'robust' criteria that pointed to the same model as the best in all cases. Simulation studies provided additional support for the practical relevance of our theoretical results, and indicated that empirical results may tend to favor an inferior model when the equivalence conditions are not satisfied.

We have also shown that a larger variance of the measurement error makes the objective-bias more likely to be relevant. This result provides an additional argument in favor of using the RV in empirical comparisons of volatility models. Yet, the RV is not a perfect measure of volatility, see Barndorff-Nielsen and Shephard (2002a), Meddahi (2002), and Andersen et al. (2003), and the recent literature on the effects of market microstructure noise is very relevant in the present context. For example, the RV is severely biased if based on intraday returns that are sampled at a high frequency, due to market microstructure noise, see e.g. Andreou and Ghysels (2002), Bandi and Russell (2004), Oomen (2002), and Zhang et al. (2005), and see Hansen and Lunde (2004b,2005b) for empirical studies that characterize the properties of market microstructure noise. So the estimators that are robust to market microstructure noise will be very useful for the evaluation and comparison of volatility models, where the subsample estimator of Zhang et al. (2005) is an attractive one, because it is consistent for the IV under certain types of market microstructure noise.

As we indicated in the introduction, this framework is applicable to other econometric problems where a proxy is substituted for a latent variable in the evaluation. Examples, of this sort include out-of-sample evaluation of forecasting models where the predicted object is a latent variable, such as forecasts of conditional quantiles and conditional densities.

Acknowledgements

We thank Sean Campbell, Andrew Patton, two anonomous referees, and the two editors, Robert Engle and Giampiero Gallo, for many valuable comments. We are responsible for all errors. Financial support from the Danish Research Agency, Grant no. 24-00-0363 and the Salomon Research Award at Brown University is gratefully acknowledged.

Appendix A. Estimation of RV_t

The standard measure of the realized variance over some interval, $[a,b]$, is usually defined by $\mathrm{RV}_{[a,b]} \equiv \sum_{i=1}^{m} y_{i,m}^2$, where $y_{i,m} \equiv p_{a+i\Delta} - p_{a+i\Delta-\Delta}$ and $\Delta = (b-a)/m$. So $y_{i,m}$, $i = 1,\ldots,m$ are intraday returns that each span a period of time with length Δ. In our empirical analysis we have applied a different method for constructing the RV, which is a Fourier method by Malliavin and Mancino (2002). This method has previously been applied by Barucci and Reno (2002a,b) and Hansen et al. (2003). A short description of the method is the following:

Let the price process be defined from a diffusion process with bounded quadratic variation, $dp(t) = \mu(t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W(t)$, where μ and σ are smooth functions and $W(t)$ is a standard Brownian motion. The integrated variance over an interval $[a,b]$ is defined by $\mathrm{IV}_{[a,b]} \equiv \int_a^b \sigma^2(t)\,\mathrm{d}t$, and the IV is typically an unbiased estimator of the conditional variance. The $\mathrm{IV}_{[a,b]}$ can be estimated by the $\mathrm{RV}_{[a,b]}$, and here we use a Fourier method for calculating the RV. The advantages of the Fourier-based RV is that it incorporates all price data and it does not require interpolation methods as is the case for the standard measure of the RV. Let $p(t)$ be observed in the interval $[t_0, t_0+1]$ at the discrete points in time, $t_1 < t_2 < \cdots < t_N$. These points in time are mapped into the interval $[0, 2\pi]$, by defining, $\tau_i = 2\pi(t_i - t_0)/(t_N - t_0)$ for $i = 1,\ldots,N$.

The Fourier method is based on the identity $1/2\pi \int_0^{2\pi} \sigma^2(t)\,\mathrm{d}t = a_0(\sigma^2)$, where

$$a_0(\sigma^2) = \lim_{m\to\infty} \frac{\pi}{2m} \sum_{k=1}^{m} (a_k^2(dp) + b_k^2(dp)) \tag{A.1}$$

and

$$a_k(dp) = \frac{p(\tau_N) - p(\tau_1)}{\pi} + \frac{1}{\pi} \sum_{i=1}^{N-1} p(\tau_i)[\cos(k\tau_i) - \cos(k\tau_{i+1})],$$

$$b_k(dp) = -\frac{1}{\pi} \sum_{i=1}^{N-1} p(\tau_i)[\sin(k\tau_i) - \sin(k\tau_{i+1})].$$

The estimate delivered by the Fourier method is given by

$$\mathrm{RV}_t \equiv \frac{\pi}{2m}\sum_{k=1}^{m}(a_k^2(dp) + b_k^2(dp)),$$

where we applied $m = 80$ in our empirical application.

Appendix B. Proofs

Proof of Lemma 1. (i) Suppose that $\mathscr{X} \succcurlyeq \mathscr{Y}$ then $\psi(\mathscr{X}) \leqslant \psi(\mathscr{Y})$. But $\tilde{\psi}(\mathscr{X}) = \psi(\mathscr{X}) - \gamma(\mathscr{X}) = \psi(\underset{\text{a}}{\mathscr{X}}) - \gamma(\mathscr{X}) + \gamma(\mathscr{Y}) - \psi(\mathscr{Y}) + \tilde{\psi}(\mathscr{Y}) = \psi(\mathscr{X}) - \psi(\mathscr{Y}) + \tilde{\psi}(\mathscr{Y}) \leqslant \tilde{\psi}(\mathscr{Y})$, and this implies $\mathscr{X} \succcurlyeq \mathscr{Y}$. The other implication of (i) is shown similarly. (ii) Suppose that $\mathscr{X} \succ \mathscr{Y}$. Then there exists $\varepsilon > 0$, such that $\psi(\mathscr{X}) + \varepsilon \leqslant \psi(\mathscr{Y})$. Similar calculations to those in the proof of (i) leads to $\hat{\psi}_n(\mathscr{X}) + \varepsilon - \delta_n(\mathscr{X}, \mathscr{Y}) \leqslant \hat{\psi}_n(\mathscr{Y})$, and since $\delta_n(\mathscr{X}, \mathscr{Y}) \overset{\text{a.s.}}{\to} 0$ it holds for almost surely that $\hat{\psi}_n(\mathscr{X}) \leqslant \hat{\psi}_n(\mathscr{Y}) - \varepsilon/2 < \hat{\psi}_n(\mathscr{Y})$, for n sufficiently large. The other implication is proven similarly. □

Proof of Theorem 2. Under Assumption (ii.a) we consider the first order Taylor expansion of L, given by $L(\tilde{\theta}_t, X_t) = L(\theta_t, X_t) + L'(\theta_t^*, X_t)\eta_t$, where θ_t^* lies between $\tilde{\theta}_t$ and θ_t. Taking expected value yields

$$E[L(\tilde{\theta}_t, X_t)] = E[L(\theta_t, X_t)] + E[L'(\theta_t^*, X_t)\eta_t].$$

Since the last term does not depend on X_t under Assumption (ii.a), it holds that

$$E[L(\tilde{\theta}_t, X_t)] - E[L(\tilde{\theta}_t, Y_t)] = E[L(\theta_t, X_t)] - E[L(\theta_t, Y_t)] \tag{A.2}$$

for all (X_t, Y_t), which shows the equivalence in this case. Under Assumption (ii.b) we consider the second order Taylor expansion of L, given by

$$L(\tilde{\theta}_t, X_t) = L(\theta_t, X_t) + L'(\theta_t, X_t)\eta_t + \tfrac{1}{2}\eta' L''(\theta_t^{**}, X_t)\eta_t,$$

where θ_t^{**} lies between $\tilde{\theta}_t$ and θ_t. Taking expected value yields

$$E[L(\tilde{\theta}_t, X_t)] = E[L(\theta_t, X_t)] + E[L'(\theta_t, X_t)\eta_t] + \tfrac{1}{2}E[\eta_t' L''(\theta_t^{**}, X_t)\eta_t],$$

where the last term does not depend on X_t, and where the second term is zero, as $E[L'(\theta_t, X_t)\eta_t|\mathscr{F}_{t-1}] = L'(\theta_t, X_t)E[\eta_t|\mathscr{F}_{t-1}] = 0$. So once $\underset{\text{a}}{\text{again}}$ we have established the identity (A.2), which shows the equivalence of $\succcurlyeq$ and $\succcurlyeq$ in this case. □

Proof of Theorem 4. The stationarity assumption implies that $E[L(\tilde{\theta}_1, X_1)] = n^{-1}\sum_{t=1}^{n} E[L(\tilde{\theta}_t, X_t)]$ and by the ergodic theorem we have that $n^{-1}\sum_{t=1}^{n} L(\tilde{\theta}_t, X_t) \overset{\text{a.s.}}{\to} E[L(\tilde{\theta}_1, X_1)]$, which completes the proof. □

Let $\tilde{\theta}_{\lambda,t} \equiv \theta_t + \lambda\eta_t$, $t = 1, \ldots, n$, and suppose that L is twice differentiable with continuous derivatives. Consider the Taylor expansion,

$$L(\tilde{\theta}_{\lambda,t}, X_t) = L(\theta_t, X_t) + L'(\theta_t, X_t)\lambda\eta_t + L''(\theta_{\lambda,t}^*, X_t)\lambda^2\eta_t^2,$$

where $\theta^{*}_{\lambda,t} \in [\theta_t, \theta_t + \lambda\eta_t]$. Suppose that $E(\eta_t|\mathscr{F}_{t-1}) \overset{\text{a.s.}}{=} 0$ and that $\text{var}(\eta_t|\mathscr{F}_{t-1}) > 0$, a.s., such that

$$\begin{aligned} E[L(\tilde{\theta}_{\lambda,t}, X_t) - L(\theta_t, X_t)] &= E[L'(\theta_t, X_t)\lambda\eta_t] + E[L''(\theta^{*}_{\lambda,t}, X_t)\lambda^2\eta_t^2] \\ &= 0 + \lambda^2 E[L''(\theta^{*}_{\lambda,t}, X_t)\eta_t^2]. \end{aligned}$$

The last term need not simplify to $\lambda^2 E[L''(\theta^{*}_{\lambda,t}, X_t)]\text{var}(\eta_t|\mathscr{F}_{t-1})$, since $\theta^{*}_{\lambda,t}$ is not $\mathscr{F}_{t-1}$-measurably (it depends on η_t).

Assumption A.1. For all $\mathscr{X} \in \mathbb{A}$, it holds that (i) $L''(\theta^{*}_{\lambda,t}, X_t) > 0$, almost surely; and (ii)

$$L''(\tilde{\theta}^{*}_{\lambda_1,t}, X_t)/L''(\tilde{\theta}^{*}_{\lambda_2,t}, X_t) \leqslant \left(\frac{\lambda_2}{\lambda_1}\right)^2,$$

almost surely, for all $0 \leqslant \lambda_1 < \lambda_2 < \infty$.

Lemma A.1. *Under Assumption* A.1, *the criterion function for the approximate preordering is given by*

$$E[L(\theta_t, X_t)] + \lambda^2 E[L''(\theta^{*}_{\lambda,t}, X_t)\eta_t^2]$$

and the measure of discrepancy is given by

$$\delta_\lambda(\mathscr{X}, \mathscr{Y}) \equiv \lambda^2 \lim_{n\to\infty} n^{-1} \sum_{t=1}^{n} E\{[L''(\theta^{*}_{\lambda,t}, X_t) - L''(\theta^{*}_{\lambda,t}, Y_t)]\eta_t^2\}. \tag{A.3}$$

Proof. The first part of the lemma follows from the calculation prior to Assumption A.1. The second part follows from the first part and the definition of $\delta_\lambda(\mathscr{X}, \mathscr{Y})$. □

Proof of Theorem 5. Consider the Taylor expansions under each of the three conditions, (i) $\varphi(\tilde{\sigma}_t^2) = \varphi(\sigma_t^2) + \dot{\varphi}\eta_t$, where $\dot{\varphi} \equiv \partial\varphi/\partial\sigma_t^2$ is a constant; (ii) $\varphi(\tilde{\sigma}_t^2) = \varphi(\sigma_t^2) + \varphi'(\sigma_t^2)\eta_t + \frac{1}{2}\ddot{\varphi}\eta_t^2$, where $\ddot{\varphi} \equiv \partial^2\varphi/\partial\sigma_t^2\partial\sigma_t^2$ is a constant; and (iii) $\varphi(\tilde{\sigma}_t^2) = \varphi(\sigma_t^2) + \varphi'(\sigma_t^2)\eta_t + 2^{-1}\varphi''(\sigma_t^2)\eta_t^2 + 6^{-1}\dddot{\varphi}\,\eta_t^3$, where $\dddot{\varphi} \equiv \partial^3\varphi/(\partial\sigma_t^2)^2$ is a constant. It now follows that for each of the three cases, $\text{cov}(\varphi(\tilde{\sigma}_t^2), \varphi(h_t)) - \text{cov}(\varphi(\sigma_t^2), \varphi(h_t))$ equals $\dot{\varphi}c_1\sigma_{hh}$, $2^{-1}\ddot{\varphi}c_2\sigma_{hh}$, and $2^{-1}\sigma_\eta^2 c_3\sigma_{hh}$, respectively. For example under condition (iii) we have

$$\begin{aligned} \text{cov}(\varphi(\tilde{\sigma}_t^2), \varphi(h_t)) &= \text{cov}(\varphi(\sigma_t^2), \varphi(h_t)) + \text{cov}(\varphi'(\sigma_t^2)\eta_t, \varphi(h_t)) \\ &\quad + 2^{-1}\text{cov}(\varphi''(\sigma_t^2)\eta_t^2, \varphi(h_t)) + 6^{-1}\text{cov}(\dddot{\varphi}\,\eta_t^3, \varphi(h_t)) \\ &= \text{cov}(\varphi(\sigma_t^2), \varphi(h_t)) + 0 + 2^{-1}\sigma_\eta^2 c_3\sigma_{hh}^2 + 0, \end{aligned}$$

where $c_3 = \text{cov}(\varphi''(\sigma_t^2), \varphi(h_t))$.

Thus $\tilde{\psi}(\mathscr{H}) = \psi(\mathscr{H}) + C$, where C equals $\dot{\varphi}^2 c_1^2$, $\ddot{\varphi}^2 c_2^2$, or $\ddot{\varphi}^2 c_3^2$ under conditions (i), (ii), and (iii), respectively. □

References

Alizadeh, S., Brandt, M., Diebold, F.X., 2002. Range-based estimation of stochastic volatility models. Journal of Finance 57, 1047–1092.

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

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2001. The distribution of exchange rate volatility. Journal of the American Statistical Association 96 (453), 42–55.

Andersen, T.G., Bollerslev, T., Meddahi, N., 2004. Analytic evaluation of volatility forecasts. International Economic Review 45, 1079–1110.

Andersen, T.G., Bollerslev, T., Diebold, F.X., 2003. Parametric and nonparametric volatility measurement. In: Aït-Sahalia, Y., Hansen, L.P. (Eds.), Handbook of Financial Econometrics, vol. I. Elsevier-North Holland, Amsterdam, forthcoming.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2003. Modeling and forecasting realized volatility. Econometrica 71 (2), 579–625.

Andreou, E., Ghysels, E., 2002. Rolling-sample volatility estimators: some new theoretical, simulation, and empirical results. Journal of Business & Economic Statistics 20 (3), 363–376.

Baillie, R.T., Bollerslev, T., Mikkelsen, H.O., 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 74, 3–30.

Bandi, F.M., Russell, J.R., 2004. Microstructure noise, realized volatility, and optimal sampling. Working paper, Graduate School of Business, The University of Chicago.

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

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

Barucci, E., Reno, R., 2002a. On measuring volatility and GARCH forecasting performance. Journal of International Financial Markets 12, 183–200.

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

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

Christoffersen, P.F., Diebold, F.X., 2002. Financial asset returns, market timing, and volatility dynamics. Manuscript, McGill University and University of Pennsylvania.

Clark, T.E., McCracken, M.W., 2001. Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85–110.

Corradi, V., Swanson, N.R., 2005. Predictive density and conditional confidence interval accuracy tests. Journal of Econometrics, forthcoming.

Corradi, V., Swanson, N.R., Olivetti, C., 2001. Predictive ability with cointegrated variables. Journal of Econometrics 104, 315–358.

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.

Elliott, G., Komunjer, I., Timmermann, A., 2005. Estimation and testing of forecast rationality under flexible loss. Review of Economic Studies, forthcoming.

Engle, R.F., 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. Econometrica 45, 987–1007.

Engle, R.F., Patton, A.J., 2001. What good is a volatility model? Quantitative Finance 1 (2), 237–245.

Giacomini, R., Komunjer, I., 2005. Evaluation and combination of conditional quantile forecasts. Journal of Business and Economic Statistics, forthcoming.

Giacomini, R., White, H., 2003. Tests of conditional predictive ability. Boston College working paper 572.

Granger, C.W.J., 1969. Prediction with a generalized cost of error function. Operational Research Quarterly 20, 199–207.

Hansen, P.R., 2001. A test for superior predictive ability. http://www.stanford.edu/people/peter.hansen.

Hansen, P.R., 2003. Asymptotic tests of composite hypotheses. http://www.stanford.edu/people/peter.hansen.

Hansen, P.R., Lunde, A., 2004a. A realized variance for the whole day based on intermittent high-frequency data. http://www.stanford.edu/people/peter.hansen.

Hansen, P.R., Lunde, A., 2004b. An unbiased measure of realized variance. http://www.stanford.edu/people/peter.hansen.

Hansen, P.R., Lunde, A., 2005a. A forecast comparison of volatility models: does anything beat a GARCH(1,1)? Journal of Applied Econometrics, forthcoming.

Hansen, P.R., Lunde, A., 2005b. Realized variance and market microstructure noise. Journal of Business and Economic Statistics, forthcoming.

Hansen, P.R., Lunde, A., Nason, J.M., 2003. Choosing the best volatility models: the model confidence set approach. Oxford Bulletin of Economics and Statistics 65, 839–861.

Hansen, P.R., Lunde, A., Nason, J.M., 2004. Model confidence sets for forecasting models. http://www.stanford.edu/people/peter.hansen.

Inoue, A., Kilian, L., 2005. On the selection of forecasting models. Journal of Econometrics, forthcoming.

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

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

Meddahi, N., 2002. A theoretical comparison between integrated and realized volatility. Journal of Applied Econometrics 17, 479–508.

Mincer, J., Zarnowitz, V., 1969. The evaluation of economic forecasts and expections. In: Mincer, J. (Ed.), Economic Forecasts and Expections. National Bureau of Economic Research, New York.

Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: a new approach. Econometrica 59, 347–370.

Oomen, R.A.C., 2002. Modelling realized variance when returns are serially correlated. Manuscript, Warwick Business School, The University of Warwick.

Pagan, A.R., Schwert, G.W., 1990. Alternative models for conditional volatility. Journal of Econometrics 45, 267–290.

Perez-Amaral, T., Gallo, G.M., White, H., 2003. A flexible tool for model building: the relevant transformation of the inputs network approach (RETINA). Oxford Bulletin of Economics and Statistics 65, 821–838.

Poon, S.-H., Granger, C., 2003. Forecasting volatility in financial markets: a review. Journal of Economic Literature 41, 478–539.

Rossi, B., 2005. Testing long-horizon predictive ability with high persistence, and the Meese-Rogoff puzzle. International Economic Review 46, 61–92.

Sin, C.-Y., White, H., 1996. Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics 71, 207–225.

Skouras, S., 2001. Decisionmetrics: a decision-based approach to econometric modelling. SFI working paper no. 01-10-64

Swanson, N.R., Zeng, T., 2001. Choosing among competing econometric forecasts: regression-based forecast combination using model selection. Journal of Forecasting 20, 425–440.

Tse, Y., 1998. The conditional heteroskedasticity of the yen–dollar exchange rate. Journal of Applied Econometrics 13 (1), 49–55.

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

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

Zakoian, J.-M., 1994. Threshold heteroskedastic models. Journal of Economic Dynamics and Control 18, 931–955.

Zhang, L., Mykland, P.A., Aït-Sahalia, Y., 2005. A tale of two time scales: determining integrated volatility with noisy high frequency data. Journal of the American Statistical Association, forthcoming.