Economics Discussion Papers
The Open-Access, Open-Assessment E-Journal

Discussion Paper 2008-18
April 23, 2008

Forecast Evaluation of Explanatory Models of Financial Return Variability

Genaro Sucarrat
Universidad Carlos III de Madrid

Abstract:

A practice that has become widespread is that of comparing forecasts of financial return variability obtained from discrete time models against high frequency estimates based on continuous time theory. In explanatory financial return variability modelling this raises several methodological and practical issues, which suggests an alternative framework is needed. The contribution of this study is twofold. First, the finite sample properties of operational and practical procedures for the forecast evaluation of explanatory discrete time models of financial return variability are studied. Second, with basis in the simulation results a simple framework is proposed and illustrated.

Paper submitted to the special issue “Using Econometrics for Assessing Economic Models” edited by Katarina Juselius.

JEL: C52, C53, F31, F37
Keywords: *Return variability forecasting; financial volatility; explanatory modelling*

Correspondence: Genaro Sucarrat, Department of Economics, Universidad Carlos III de Madrid (Spain). Email: gsucarra@eco.uc3m.es.

I am indebted to Álvaro Escribano and people at the International Symposium of Forecasting in June 2007 for useful comments, questions and suggestions. I am also indebted to Peter R. Hansen for kindly providing me the source code of the SPA package used in some of the simulations. Errors and interpretations being my own applies of course. This research was supported by a Marie Curie Intra-European Fellowship within the 6th European Community Framework Programme.

1 Introduction

Explanatory models of financial return variability can be useful in a wide range of situations. In risk management explanatory models are useful in stress-testing, event analysis, conditional forecasting and counterfactual analysis. In asset pricing explanatory models provide a more detailed way of describing the price variation of the underlying asset, and enables asset pricing conditional on the values of impact variables. In policymaking explanatory models can shed light on the impact of a change in the interest rate, of currency market interventions, of changes in regulatory regime, of changes in liquidity, and so on. Forecast comparison plays an informative role in the evaluation of explanatory models intended for any of these purposes.

A practice that has become widely endorsed is that of comparing forecasts of financial return variability obtained from discrete time models against high frequency estimates based on continuous time theory, see amongst others Zhou (1996), Taylor and Xu (1997), Andersen and Bollerslev (1998), Andersen et al. (1999), Meddahi (2002), Andersen et al. (2003), Hansen and Lunde (2006), and Andersen et al. (2006). In particular, numerous studies investigate and/or use realised volatility—the sum of intra-period squared returns—either as comparison benchmark or as a measure of "true" volatility. The main motivation for this is that the high frequency estimate is believed to provide a more efficient estimate of volatility. In explanatory financial return modelling, however, this raises several methodological and practical issues. First, in empirical modelling the error term and the standardised residual are derived in the sense that their properties depend on the functional form and explanatory information in the conditional mean and conditional variance specifications. Since explanatory information typically is less available at high frequencies, explanatory low-frequency models can produce substantially better estimates of return variability than high frequency models due to differing information sets. So procedures for evaluating them against each other without treating either as more basic is needed. Second, since time is needed for an event to bring about another event, explanatory variables are likely to account for a decreasing portion of return variability as the time increment goes to zero. Indeed, a large part of modern finance theory is based on the idea that private information disseminates sequentially and aggregates temporally. When the time increment goes to zero, then the existence of such an effect is questionable. An example of special interest in this context is the effect of order imbalances or order flow, since it has been shown that order flow can explain a substantial portion of return variation, see amongst others Blume et al. (1989), Lee and Ready (1991), Hasbrouck (1991), Chordia et al. (2002), Evans and Lyons (2002), Engle and Patton (2004), Dunne et al. (2005), Escribano and Pascual (2006), and Moberg and Sucarrat (2007). So even when explanatory information is available at high frequencies, explanatory modelling at lower frequencies can produce different results due to aggregation issues. Third, it is well-known that market microstructure issues affect the precision of high-frequency estimates, see amongst others Meddahi (2002),Barndorff-Nielsen and Shephard (2002b) Aït-Sahalia and Mykland (2003), Andersen et al. (2005), and Aït-Sahalia et al. (2005). Hence, it is not given that high-frequency estimators perform better than low-frequency estimates, and procedures for choosing between them is needed without *a priori* treating either as more fundamental. Fourth, the right combination of information and functional form in the conditional mean can result in homoscedastic errors. For example, an explicit aim of the General-to-Specific (GETS) methodology is to specify the conditional mean such that the errors become homoscedastic, since heteroscedasticity frequently is an indication of missing variables and/or structural breaks, see Gilbert (1990), Mizon (1995) or Campos et al. (2005) for overviews of the GETS methodology. It is inappropriate to compare the constant volatility estimate implied by homoscedasticity with a time-varying high-frequency estimate, so alternative comparison procedures are needed. Finally, comparing the estimates from a discrete time explanatory model with high-frequency estimates based on continuous time theory constitutes a probabilistic restriction, since discrete models are compatible and can be derived from more than one continuous time structure, see Sucarrat (2007, section 4.2) for a discussion.

Together these methodological and practical issues suggest that an alternative framework is needed when comparing explanatory models' forecasts of financial return variability. The contribution of this study is twofold. First, the finite sample properties of operational and practical procedures for the evaluation of explanatory discrete time models of financial return variability are investigated through a simulation study, where return variability is defined as squared return. The main advantage of using an observable yardstick or reference point like squared return is that forecast accuracy can then be measured with respect to an objectively given entity rather than a magnitude that is entirely determined by the assumptions of the postulated continuous time model. Appropriate understanding in finite samples is crucial since explanatory data typically is available at lower frequencies only, say, daily, weekly, monthly, etc. Second, with basis in the simulation results a simple framework for the evaluation of explanatory models of return variability is proposed and illustrated. In the illustration the explanatory model of financial return is the best *ex post* forecaster of variability—including better than realised volatility, whereas the results of the *ex ante* evaluation do not suggest that any candidate is better than a constant forecast of return variability.

The rest of the paper is divided into four sections. The next section provides a more detailed characterisation of some of the methodological and practical issues that arise in the evaluation of explanatory models of return variability. Section three contains the simulation study and the simple framework that is suggested with basis in the simulation results. Section four illustrates the use of the framework applied to *ex post* and *ex ante* out-of-sample forecast evaluation, using data that are particularly prone to the methodological and practical issues that arise when evaluating explanatory models of financial return variability against high frequency estimates based on continuous time theory. Finally, section five concludes and gives suggestions for further research.

2 Explanatory modelling of financial return variability

The purpose of this section is to provide a more detailed characterisation of the methodological and practical issues that arise in forecast evaluation of explanatory models of financial return variability. The section consists of two subsections. The first subsection contains a discussion of explanatory discrete time models as derived entities. This discussion is needed in order to understand why volatility in explanatory models is not simply a latent variable to be estimated, but rather an entity whose properties depend on functional form and the explanatory power of the information in the conditional mean and variance specifications. The second subsection discusses in more detail some of the issues that arise in evaluating discrete time model estimates against estimates based on continuous time theory.

2.1 Discrete time models as derived entities

Econometric models are simplified and partial representations of a highly complex and evolving social reality, and the probabilistic study of their relation belongs to reduction theory, see amongst others Hendry and Richard (1982), Florens et al. (1990), Hendry (1995, chapter 9) and Sucarrat (2007).[1] A key distinction in reduction theory is that between the model that governs the reality on the one hand and simplifications of it on the other, and the objective of reduction theory is to study to what extent important information is lost by representing the former by means of the latter. A well-known example of a model that governs reality is the Data Generating Process (DGP) as defined in David F. Hendry's (1995, chapter 9) reduction theory. On the other hand, a simple example of a simplification of the DGP is the linear model $r_t = b_0 + b_1 x_t + e_t$, where r_t and x_t are variables, b_0 and b_1 are coefficient values, and e_t is the error. A shortcoming with Hendry's framework is that it cannot provide reduction analysis on the relation between continuous and discrete time models, since the framework is entirely couched in terms of discrete time variables. However, the non-restrictive modifications to the initial probability space in Hendry's theory proposed in Sucarrat (2007) enables reduction analysis on the relation between continuous and discrete time models.

A key implication of reduction theory is that the properties of the empirical model are the result of its specification subject to the model that governs reality. To see the implication of this in a volatility context consider first the implication for models of financial returns. For simplicity of discussion but with no loss of generality, let us henceforth assume that there is no measurement error in any of the variables of the DGP, and that the variables of the DGP correspond to that of the theory mechanism. (It should be stressed that volatility *is not* a variable in

[1] Reduction theory plays an important role in the General to Specific (GETS) methodology, since the methodology can be viewed as an attempt to mimic reduction theory.

the theory mechanism nor in the DGP, since the theory mechanism and the DGP are entities whose properties are independent of how we represent them my means of models, see Sucarrat 2007 for further discussion.) In other words, the DGP is assumed equal to the theory mechanism. Let the density $f(r_t, \mathbf{x}_t, \mathbf{y}_t)$ denote the DGP of r_t, $\mathbf{x}_t$ and $\mathbf{y}_t$, where r_t is a financial return, and where $\mathbf{x}_t$ and $\mathbf{y}_t$ are vectors of "explanatory" variables. In addition to other contemporaneous and/or lagged variables the vectors $\mathbf{x}_t$ and $\mathbf{y}_t$ may contain lags of r_t and/or transformations of lags of r_t in empirical applications. Suppose the discrete time representation

$$r_t = g(\mathbf{x}_t, \mathbf{b}) + e_t,$$

is an empirical model of the conditional DGP given by $f(r_t|\mathbf{x}_t)$, where $\mathbf{b}$ is a parameter vector, $g(\mathbf{x}_t, \mathbf{b})$ is equal to the conditional mean $E(r_t|\mathbf{x}_t)$ and e_t is the error term. The error term e_t is defined as $r_t - g(\mathbf{x}_t, \mathbf{b})$, and its properties are therefore derived or "designed" in the sense that they are a result of how $g(\mathbf{x}_t, \mathbf{b})$ is specified subject to the conditional DGP given by $f(r_t|\mathbf{x}_t)$.[2] In particular, the better $g(\mathbf{x}_t, \mathbf{b})$ is specified and the more explanatory power carried by $\mathbf{x}_t$, the smaller e_t is likely to be in absolute value.

Consider now the discrete time model

$$r_t = g(\mathbf{x}_t, \mathbf{b}) + e_t, \qquad e_t = \sigma_t z_t, \tag{1}$$

$$\sigma_t^2 = h(\mathbf{y}_t, \mathbf{c}), \tag{2}$$

where $\mathbf{c}$ is a vector of parameters, and where σ_t^2 is discrete time volatility and equal to the conditional variance $Var(r_t|\mathbf{x}_t, \mathbf{y}_t)$. The term $g(\mathbf{x}_t, \mathbf{b})$ is now equal to the conditional mean $E(r_t|\mathbf{x}_t, \mathbf{y}_t)$, and the standardised residual z_t is defined as $[r_t - g(\mathbf{x}_t, \mathbf{b})]/\sigma_t$. The properties of z_t are therefore determined by the specification of $g(\mathbf{x}_t, \mathbf{b})$ and $h(\mathbf{y}_t, \mathbf{c})$ subject to the conditional DGP given by $f(r_t|\mathbf{x}_t, \mathbf{y}_t)$. In particular, the better $g(\mathbf{x}_t, \mathbf{b})$ and $h(\mathbf{y}_t, \mathbf{c})$ explain the variation in r_t and e_t^2, respectively, the smaller z_t is likely to be in absolute value.

2.2 Continuous vs. discrete models

If explanatory models are derived entities that depend on the specification of the conditional mean and variance subject to the DGP, then volatility is not just a *given* unobservable magnitude as suggested by some scholars. On the contrary, the value and characteristics of volatility depends on the conditional mean and

[2] There are at least two possible sources of information loss in modelling r_t by means of $g(\mathbf{x}_t, \mathbf{b})$. First, the variables $\mathbf{y}_t$ have been marginalised, so one may ask how well $f(r_t|\mathbf{x}_t)$ approximates $f(r_t|\mathbf{x}_t, \mathbf{y}_t)$. Second, there is the question of how well the distribution of $g(\mathbf{x}_t, \mathbf{b}) + e_t$ approximates $f(r_t|\mathbf{x}_t)$, see Hendry (1995, chapter 9) for a more detailed discussion.

variance specifications, and the more so the better the explanatory variables explain the variation in return and in the squared error. Accordingly, comparing volatility estimates from an explanatory discrete time model against high-frequency estimates that are based on continuous time theory can lead to highly misleading results. For the purpose of a more specific discussion, consider as an example of a general class of continuous time models the semi-martingale

$$r(t) = A(t) + M(t), \quad t \in [0, T], \tag{3}$$

where $r(t) = p(t) - p(t-1)$ is the price increment from $t-1$ to t, $A(t)$ is a locally integrable and predictable process of finite variation, and $M(t)$ is a local martingale, see Andersen et al. (2001) and Andersen et al. (2003). Some continuous time models that are contained in this formulation are Itô, jump and jump-diffusion processes. For example, by setting $A(t)$ equal to $\int_{t-1}^{t} \mu(s)ds$ and $M(t)$ equal to $\int_{t-1}^{t} \sigma(s)W(s)ds$, where $\{\mu\}$ and $\{\sigma\}$ are continuous processes, and where $\{W\}$ is a standard Wiener process, we obtain the Itô process

$$r(t) = \int_{t-1}^{t} \mu(s)ds + \int_{t-1}^{t} \sigma(s)W(s)ds. \tag{4}$$

In this particular case the quadratic variation $\int_{t-1}^{t} \sigma(s)^2 ds$ serves as the counterpart of discrete time volatility σ_t^2 as defined in the discrete time model (1)-(2) above, and a common estimator of quadratic variation is realised volatility, that is, the sum of equi-distant intraperiod squared returns.

Evaluating volatility estimates obtained from the discrete time model (1)-(2) against estimates obtained based on (say) (3) or (4) raises several methodological and practical issues which were alluded to in the introduction. Here four of these issues are discussed in more detail. First, although return data may be available at high frequencies (say, intradaily) this is not necessarily the case for explanatory data. Suppose for example that order flow data is available at lower frequencies but not at higher frequencies. That means the term $\int_{t-1}^{t} \mu(s)ds$ is likely to explain a very small fraction of the total variation in $r(t)$ when estimated on high-frequency data. When estimated on lower frequency data by contrast the term might account for a substantial fraction of the total return variation. A similar argument applies of course to values of $\int_{t-1}^{t} \sigma(s)ds$. In the absence of explanatory information its value is entirely determined by the assumption that $\int_{t-1}^{t} \mu(s)ds$ is equal to or approximately zero, and by the assumptions regarding the process $\{W(t)\}$. The lower frequency model by contrast may produce substantially different values of $\int_{t-1}^{t} \sigma(s)ds$ due to explanatory information in either or both $\int_{t-1}^{t} \mu(s)ds$ and $\int_{t-1}^{t} \sigma(s)ds$. As a consequence, the properties of $\int_{t-1}^{t} W(s)ds$ can be substantially different from the high frequency model. All this is not surprising since the two approaches use different information sets, but how then should we evaluate them against each other without treating either as more basic? Two solutions that suggest themselves naturally is to evaluate the loss associated with prediction errors of return variability, say, r_t^2, and to evaluate the total fit in terms of the sample kurtosis of the standardised residual. The latter is of interest because it is analogous to the standard error of a homoscedastic regression. The usefulness of both of these strategies will be studied in the next section. Second, for economic reasons $A(t)$ and the explanatory component of $M(t)$—for example $\int_{t-1}^{t} \sigma(s)ds$ in (4)—are likely to account for a decreasing portion of the variation in returns as the time increment decreases, since time is needed for an event—or as is typically the case, a combination of events—to bring about another event. The economic reasoning underlying Evans and Lyons' (2002) order flow measure, for example, is that private information disseminates sequentially and aggregates temporally, so that time is needed for it to have an effect. In other words, even if explanatory intra period high-frequency data is available, an inter period low frequency model of variability may perform better. Third, it is well-known that microstructure issues, measurement errors and finite sample issues affect—possibly in substantial ways—the accuracy of high-frequency intra period estimates, see amongst others Meddahi (2002), Barndorff-Nielsen and Shephard (2002a), Barndorff-Nielsen and Shephard (2002b) Aït-Sahalia and Mykland (2003), Andersen et al. (2005), Aït-Sahalia et al. (2005) and Aït-Sahalia (2006). The nature and magnitude of the estimation error is never known with exact precision, so one cannot be sure that the chosen error-adjusting method appropriately corrects for the error. Also, it is not given that the chosen error-adjusting method performs better than low-frequency estimates, possibly with explanatory information in the conditional mean and variance or both. Accordingly, procedures that enable us to evaluate the estimates against each other are needed. Finally, whenever it is assumed that the discrete time model can be derived from the continuous time model—as assumed in Andersen and Bollerslev (1998)—then a probabilistic restriction is imposed. In other words, contrary to a common misperception a continuous time model does not nest a discrete time model if the latter can be derived from the former. The reason for this is that discrete time models are potentially compatible (and can thus be derived from) more than one continuous time structure. In terms of the concepts and terminology proposed in Sucarrat (2007, see section 4.2. in particular), if $\mathcal{A} = \{A_1, A_2, \ldots\}$ are the sets of possible worlds in which the discrete time model (1)-(2) is "true", and if $\mathcal{B} = \{B_1, B_2, \ldots\}$ are the sets of possible worlds in which the continuous time model (3) is true, then the probabilities of (1)-(2) and (3),respectively, being true are $P(\bigcup_{i=1}^{\infty} A_i)$ and $P(\bigcup_{j=1}^{\infty} B_j)$, respectively. Furthermore, the probability of both (1)-(2) *and* (3) being true jointly, which is effectively the assumption upon which evaluation of discrete time estimates against continuous time estimates is based, is $P[(\bigcup_{i=1}^{\infty} A_i) \cap (\bigcup_{j=1}^{\infty} B_j)] \leq P(\bigcup_{i=1}^{\infty} A_i)$. In words, the probability that both the discrete time model (1)-(2) and the continuous time model (3) are true is always equal to or smaller than the probability that only the discrete time model (1)-(2) is true. Another way to put this is that, in a probabilistic sense, continuous time models do *not* nest the discrete time models that can be derived from the former unless the joint probability of both being true is equal to 1.

3 Forecast evaluation of explanatory models of financial return variability

The main consequence of the discussion in the previous section is that evaluation of discrete models of volatility against high-frequency estimates of continuous time analogs can be misleading, in particular when the explanatory information in the mean and/or in the variance specifications have notable explanatory power. Accordingly, we need comparison procedures that enable us to evaluate discrete time and continuous time estimates against each other without treating either as more basic *a priori*. This section evaluates such accuracy measures by means of a simulation study that pays particular attention to two questions: (1) What the most appropriate loss function is, and (2) What the most appropriate out-of-sample forecast test is. A substantive number of studies have contributed directly or indirectly to the understanding of these questions within the paradigm of volatility being *given* (as opposed to *determined* by the explanatory information included), see amongst others Andersen and Bollerslev (1998), Meddahi (2002), Andersen et al. (2005), Hansen and Lunde (2005, 2006), and Patton (2007). Here, by contrast, the aim is to shed light on variability model evaluation within the paradigm of volatility *not* being given, but a result of the information included in the mean and variance specifications. As a consequence, loss will be conceived in terms of the forecast error of squared returns, since squared returns is a measure of the total variation in return or price variability, and in terms of the kurtosis of standardised residuals.[3] The sample kurtosis of the standardised residuals is of interest because it is analogous to the standard error of a homoscedastic regression. In other words, the sample kurtosis of the standardised residuals can be seen as a goodness-of-fit measure of how much the mean and variance specifications explain jointly of the total variation in return.

The section consists of six subsections. The first subsection describes and motivates the simulation setup. Next, subsection two motivates and describes the comparison models. Subsection three sheds light on what the most appropriate loss function is, whereas subsections four and five studies the appropriateness of some common out-of-sample forecast tests. Finally, with basis in the simulation results, subsection six outlines a simple framework for out-of-sample return variability comparison, which is illustrated in use in the next section.

3.1 Simulation setup

The data-generating process (DGP) of the simulations is given by

[3] Another natural candidate is the loss associated with the forecast error of absolute returns. However, the forecast error of squared return is more common in the literature and computationally easier to work with.

$$
\begin{aligned}
r_t &= bx_t + e_t, \quad e_t = \sigma_t z_t, \quad x_t \sim IIN(0,1), \quad z_t \sim IIN(0,1), \\
\sigma_t^2 &= \omega + \alpha e_{t-1}^2 + \beta \sigma_{t-1}^2 + cy_t, \quad y_t \sim IID, \quad y_t \in \{0,1\} \text{ with } P(1) = p,
\end{aligned}
\tag{5}
$$

for $t = 1, \ldots, T$, where x_t, z_t and y_t are mutually independent for all t. The term bx_t is the explained portion of conditional first moment return variation, and in econometric practice the explanatory power can be due to (say) contemporaneous and/or lagged money market variables ("interest rates"), stock market variables, order flow variables, news variables, and so on. For simplicity reasons the volatility persistence term in the variance specification σ_t^2 is specified as a GARCH(1,1) structure $\alpha e_{t-1}^2 + \beta \sigma_{t-1}^2$, where $0 < \alpha + \beta < 1$ (the closer $\alpha + \beta$ is to 1, the greater volatility persistence, and $\alpha + \beta \geq 1$ implies non-stationarity). However, in principle the persistence term can be replaced by explanatory variables that account for volatility persistence, for example volume and/or other liquidity variables or other more complicated persistence structures. The last term cy_t in the variance specification is a "jumpy" or non-persistent component. The value c is a non-negative scalar and $\{y_t\}$ is a two-valued IID process with probabilities $P(1) = p$ and $P(0) = 1 - p$, respectively. In empirical practice the explanatory power of the jump component can also be due to explanatory variables, for example contemporaneous and/or lagged news and/or unexpected events, shocks, and so on, and again the specification of the jump component is due to simplicity reasons.

Letting $\mathcal{I}_t$ stand for the contemporaneous and past conditioning variables $\{x_t, y_t, x_{t-1}, e_{t-1}, \sigma_{t-1}, y_{t-1}, \ldots\}$, then the conditional mean $E(r_t|\mathcal{I}_t)$ of the simulation DGP (5) is bx_t, the conditional variance (volatility) $Var(r_t|\mathcal{I}_t)$ is σ_t^2, the conditional variability $E(r_t^2|\mathcal{I}_t)$ is $(bx_t)^2 + \sigma_t^2$, and the standardised residual z_t is $(r_t - bx_t)/\sigma_t$. A measure of the total variation in r_t is given by variability r_t^2, and two definitions of the explained portion of variability are conditional variability $(bx_t)^2 + \sigma_t^2$ and unconditional variability $b^2 + \frac{\omega}{1-\alpha-\beta} + \frac{cp}{1-\alpha-\beta}$, respectively. Unconditional variability is thus made up of three separate terms: An explanatory term b^2 stemming from the conditional mean, a term $\frac{\omega}{1-\alpha-\beta}$ that is due to volatility persistence and a term $\frac{cp}{1-\alpha-\beta}$ that is due to the jump component. In order to compare the impact of each of the three terms a benchmark simulation DGP—a reference point—will be specified such that, unconditionally, each of the three terms account for an equal portion of the total explained unconditional variability. In other words, in the benchmark DGP the restriction $b^2 = \frac{\omega}{1-\alpha-\beta} = \frac{cp}{1-\alpha-\beta}$ is imposed on the choice of the parameter values. Moreover, because financial returns are commonly found to be volatility persistent, and since it is of interest to study the impact of high persistence, α and β are set to 0.1 and 0.8, respectively. In order to further calibrate the simulation set-up such that it becomes realistic, ω is set to 0.02. This implies that the term $(\frac{\omega}{1-\alpha-\beta})^{1/2} = 5^{-1/2} \approx 0.45$, which is virtually identical to the sample standard deviation of interdaily USD/EUR returns in table 1. In other words, in the case where $b = 0$ and $c = 0$ in the benchmark values the simulation DGP produces an unconditional standard deviation of returns equal to the empirical estimate of the daily standard deviation of USD/EUR returns in the period 30 September 2005 - 4 January 2008. The jump probability p is set to 0.1, which means there is a jump every tenth observation on average, and consequently $c = 2$ and $b = 5^{-1/2}$. Writing $\mathbf{a} = (b, \omega, \alpha, \beta, c, p)$ for notational convenience we therefore have that the benchmark simulation DGP is given by (5) with $\mathbf{a} = (5^{-1/2}, 0.02, 0.1, 0.8, 2, 0.1)$.

Table 3 contains descriptive statistics of simulated returns for different values of $\mathbf{a}$, and compares with tables 1 and 2 which contain descriptive statistics of the returns of four selected daily and weekly exchange rates from 30 September 2005 to 4 January 2008. The data and the sample choice are partly due to the empirical application in the next section. For $\mathbf{a}_1$, which is equal to the benchmark values $(5^{-1/2}, 0.02, 0.1, 0.8, 2, 0.1)$, the simulated standard error is 0.77, which is higher than than the four daily standard errors and about the same as the lower of the two weekly standard errors. Removing the jump term, the persistence term and the explanatory term—this gives $\mathbf{a}_5$—reduces the standard error of simulated returns to 0.45. This is equal to the daily standard error of the USD/EUR exchange rate, and slightly higher than the GBP/EUR and NOK/EUR exchange rates. The highest kurtosis among the simulated returns is produced by $\mathbf{a}_1$ and is equal to 3.09. This is relatively low since only weekly USD/EUR and GBR/EUR returns are lower among the eight empirical estimates. Four of the remaining six kurtosis values of the empirical exchange rate returns range from 3.571 to 3.941, whereas the kurtosis of YEN/EUR returns are as high as 4.918 and 7.047 in the daily and weekly cases, respectively. The high kurtosis of YEN/EUR returns are due to large (in absolute value) returns, which to some extent are clustered (less so in the weekly case). This suggests that setting the jump size c equal to 2—as in $\mathbf{a}_1$—is relatively low compared with empirical returns, or at least for YEN/EUR returns from September 2005 to January 2008. The fourth row in table 3 contains the coefficient of multiple correlation R^2 of the OLS estimated regression $r_{lt} = \hat{\gamma}_0 + \hat{\gamma}_1 x_t + \hat{e}_{lt}$, and shows that the jump term have a large impact on the explanatory power of x_t, and that the persistence term does not have an impact on the explanatory power. With no jump term R^2 is as high as 50%, regardless of whether the persistence term is included or not. Including the jump term, however, reduces R^2 to 35%. The fifth and final row in table 3 contains the R^2 of the OLS estimated regression $r^2_{lt} = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{r}^2_{lt} + \hat{e}_{lt}$, where $\hat{r}^2_{lt}$ is conditional variability of returns under $\mathbf{a}_l$. It is commonly found that these socalled Mincer and Zarnowitz (1969) regressions of r^2_t on forecasts of variability exhibit very low explanatory power in terms of the R^2, see Andersen and Bollerslev (1998). The simulations suggest that the low explanatory power (in population terms) remains low even for $\mathbf{a}_3$, where the conditional mean accounts for as much as 50% of return variation, and where there is no heteroscedasticity in the errors of the simulation DGPs.

3.2 Comparison models

Four models will be studied and compared given (5) as the DGP. The first of the four models is intended to mimic the situation where one fits a model that includes all three explanatory components of the explained variation in returns. Specifically, the forecasts of r_t and σ_t^2 for model 1 are given by

$$\hat{r}_{1t} = bx_t, \qquad \hat{\sigma}_{1t}^2 = \omega + \alpha e_{1t-1}^2 + \beta\sigma_{t-1}^2 + cy_t \tag{6}$$

where $e_{1t} = r_t - bx_t$, and the model's standardised residual at t is given by $\hat{z}_{1t} = (r_t - \hat{r}_{1t})/\hat{\sigma}_{1t}$. Accordingly, by construction $\hat{z}_{1t} = z_t$ in the simulations. The second model is intended to mimic the situation where one fits a model that only includes the persistence and jump terms, which means the conditional mean is set to zero. Specifically, forecasts of r_t and σ_t^2 for model 2 are given by

$$\hat{r}_{2t} = 0, \qquad \hat{\sigma}_{2t}^2 = [\omega + b^2(1-\alpha-\beta)] + \alpha e_{2t-1}^2 + \beta\sigma_{t-1}^2 + cy_t \tag{7}$$

where $e_{2t} = r_t$, and the model's standardised residual $\hat{z}_{2t}$ is defined as $r_t/\hat{\sigma}_{2t}$. The "augmented" constant $[\omega + b^2(1-\alpha-\beta)]$ in the variance specification is intended to adjust for the absence of bx_t in the mean specification, and ensures that the unconditional variability $E(\hat{r}_{2t}^2)$ of model 2 is equal to the true unconditional variability $E(r_t^2)$ in the limit. For simplicity the DGP values of the parameters α, β and c are used, but in econometric practice they would be determined by the choice of estimation method, say, Quasi Maximum Likelihood (QML) estimation. The third model is intended to mimic the situation where one fits a model that only includes the persistence term out of the three explanatory components. In other words, the conditional mean is set to zero and there is no jump term in the conditional variance. Specifically, forecasts of r_t and σ_t^2 for model 3 are given by

$$\hat{r}_{3t} = 0, \qquad \hat{\sigma}_{3t}^2 = [\omega + b^2(1-\alpha-\beta) + cp] + \alpha e_{3t-1}^2 + \beta\sigma_{t-1}^2 \tag{8}$$

where $e_{3t} = r_t$, and the model's standardised residual $\hat{z}_{3t}$ is defined as $r_t/\hat{\sigma}_{3t}$. Here the augmented constant is specified as $[\omega + b^2(1-\alpha-\beta) + cp]$ in order to adjust for the zero mean specification and the absence of the jump term in the variance specification. Again, the motivation is to ensure that the unconditional variability of model 3 is equal to the true unconditional variability in the limit. Finally, the fourth model is intended to mimic the situation where one uses the sample variance as an estimate of variability. Specifically, forecasts of r_t and σ_t^2 for model 4 are given by

$$\hat{r}_{4t} = 0, \qquad \hat{\sigma}_{4t}^2 = \text{simulated sample variance of } r_t, \tag{9}$$

where the simulated sample variance is that obtained from the simulations reported in table 3. In other words, for the benchmark simulations $\mathbf{a}_1$ its value is approximately $(0.94)^2 \approx 0.88$. The standardised residual $\hat{z}_{4t}$ is defined as $r_t/\hat{\sigma}_{4t}$.

3.3 What is the most appropriate loss function?

By construction model 1 accounts for a greater proportion of explained conditional variability than model 2, model 2 accounts for a greater proportion of explained conditional variability than model 3, and model 3 accounts for a greater proportion of explained conditional variability than model 4. But to what extent are loss functions capable of reproducing this ranking? This is the question to be addressed in this section. Numerous loss functions have been used, studied and suggested in the volatility evaluation literature within the paradigm of volatility being given, see Patton (2007) for a survey, only four will be compared here. The first of the loss functions that will be studied is mean squared error (MSE) of variability forecasts, and arguably MSE is the most commonly used loss function in econometric volatility evaluation. The MSE of model m is given by

$$MSE_m = \frac{1}{T}\sum_{t=1}^{T}(r_t^2 - \hat{r}_{mt}^2 - \hat{\sigma}_{mt}^2)^2. \tag{10}$$

The lower the MSE_m, the greater proportion of variability r_t^2 is on average explained by model m. A possible shortcoming with the MSE measure is that it is biased towards rejecting models unless they explain a substantial proportion of variability. This motivates the second measure, the mean absolute error (MAE). The MAE of model m is given by

$$MAE_m = \frac{1}{T}\sum_{t=1}^{T}|r_t^2 - \hat{r}_{mt}^2 - \hat{\sigma}_{mt}^2|. \tag{11}$$

The lower the MAE_m, the greater proportion of variability is explained by model m. The third type of loss function that will be studied is the sample kurtosis of the standardised residual. The sample kurtosis of model m is given by

$$K_m = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{r_t - \hat{r}_{mt}}{\hat{\sigma}_{mt}}\right)^4. \tag{12}$$

The motivation for the sample kurtosis of the standardised residual as a loss function is that it constitutes a goodness-of-fit measure in a time-varying volatility context, analogous to the standard error of a regression in a constant-volatility context. The better the model jointly accounts for the time-varying variation in return and in the error term, the lower the kurtosis. Finally, the fourth type of loss function that will be studied is the multiple correlation coefficient R^2 of regressions of the type

$$r_t^2 = a + b(\hat{r}_{mt}^2 + \hat{\sigma}_{mt}^2) + u_{mt}, \quad t = 1, \ldots, T. \tag{13}$$

These regressions are commonly referred to as Mincer-Zarnowitz regressions after Mincer and Zarnowitz (1969), and have proved useful in the forecast evaluation of a range of different economic and financial series.

Table 4 contains the simulated probabilities of obtaining the correct ranking of all four models for the benchmark values $\mathbf{a}_1$. Ideally, the probability of providing a correct ranking should increase with sample, and the only loss function that does not produce this characteristic is the residual kurtosis K: Its probability is virtually zero at all sample sizes, and decreases to zero as the sample size increases. For MSE and R^2 by contrast the probability increase with sample size. For MSE the probability increases from 28% when $T = 25$ to 96% when $T = 1000$, whereas for R^2 the probability increases from 33% when $T = 25$ to 99% when $T = 1000$. In other words, the R^2 of Mincer-Zarnowitz regressions seems more likely than MSE in providing the correct ranking in finite samples. It should be noted however that a possible reason for this is that the constant model—which in the simulation by construction is the worst—always produces an R^2 of zero. This suggests that MSE is preferable to R^2. The MAE seems to be increasing in probability until $T = 500$ where the probability is 62%, but then for $T = 1000$ the probability drops 1 percentage point to 61%. Closer inspection of the simulation results reveals that the source of this is the constant model, see table 8. The probabilities of correctly ranking the other models always increase with sample size, but not for model 4. First the probability increases to 97% until $T = 100$, then it starts to decrease at some point before $T = 500$ while ending up at 84% when $T = 1000$. Another result that is clear from the simulations and which is of practical interest is that in small samples, say, (approximately) when $T < 100$, the MAE is considerably more likely to provide a correct ranking than MSE and R^2, and the magnitude is sufficiently high to be of practical use. For example, for $T = 25$ the probabilities of MSE, MAE and R^2 are 28%, 44% and 33%, respectively, for $T = 50$ the probabilities are 39%, 49% and 47%, whereas for $T = 100$ the probabilities are 52%, 57% and 65%. Additional simulations (not reported) with different parameter values predictably suggest that the probabilities fall as the difference between the models is reduced, and that the probabilities fall as the values of the parameters b, α, β and c are reduced. However, the property that probabilities increase with sample size when MSE, MAE and R^2 are used is retained (an exception is model 4 when MAE is used in large samples, see discussion below).

Tables 5 to 8 contain the simulated probabilities for each of the model's rank when the DGP is given by (5) and the benchmark values $\mathbf{a}_1$. There are at least three characteristics of interest that emerge from the results. First, the only loss functions that always (that is, for all four models) yield highest probabilities for the correct ranking regardless of sample size are MSE and MAE. The R^2 yields highest probability for the correct ranking most of the time, but not when $T = 25$ for models 2 and 3. For these models the R^2 have a small probability of 2% to incorrectly rank model 3 in front of model 2. Kurtosis by contrast performs miserably since its highest probability rarely corresponds to the correct ranking of the model in question. A second characteristic of the results, which is in favour of the MAE in samples smaller than (approximately) 100 observations, is that it in such situations MAE is more likely than both MSE and R^2 to correctly rank each model regardless of the others' rank. For example, ranking according to MAE when $T = 25$ then the respective probabilities for models 1 to 4 are as high as 70%, 56%, 51% and 91%. By contrast, the probabilities for MSE when $T = 25$ are 67%, 42%, 42% and 69%. The probabilities of R^2 are generally comparable to or slightly higher than those of MSE. However, a possible reason is that the fact that model 4 by construction is the worst, and that R^2 always ranks model 4 last. A further implication of all this is that the probability of correctly ranking a single model correctly regardless of the correctness of the other models' rank can be substantially higher than the probability of correctly ranking all four models simultaneously. This is useful when one is interested in evaluating a certain model against a set of comparison models rather than obtaining the correct ranking between all the models. Finally, a third finding that is clear from the results is that increasing the sample size increases the probability of ranking each model correctly regardless of the others' ranks if MSE and R^2 are used, but not always when MAE is used. For model 4 the probability of correctly ranking it last falls from 90% to 84% when T increases from 500 to 1000. As noted above, additional simulations predictably suggest that the probabilities fall as the difference between the models is reduced, and that the probabilities fall when the size of the parameter values is reduced. However, the property that probabilities increase with sample size when MSE, MAE and R^2 are used is retained also for each model's rank regardless of the others' rank.

3.4 Multiple comparison tests

The loss functions MSE, MAE, Kurtosis and the R^2 of Mincer-Zarnowitz regressions can provide rankings of the variability forecasts, but the measures alone do not give any information regarding the statistical significance of the forecast properties. A common econometric evaluation strategy is that of assessing whether the loss associated with the forecast errors of one or several models is significantly smaller than the loss associated with the forecast errors of a benchmark model. Three examples of tests that can be used for this purpose are the modified version of Diebold and Mariano's (1995) comparative forecast accuracy test (MDM), see Harvey et al. (1997), White's (2000) socalled "reality check" (RC) and Hansen's (2005) test for superior predictive ability (SPA). If $g(r_t, \hat{r}_{mt}, \hat{\sigma}^2_{mt})$ denotes the loss associated with the predictions of model m at t, and if $g(r_t, \hat{r}_t, \hat{\sigma}^2_t)$ denotes the loss associated with the benchmark model at t, then the MDM test provides a simple and flexible way of testing the null of the benchmark yielding less or equal loss, that is, $E[g(r_t, \hat{r}_t, \hat{\sigma}^2_t)] \leq E[g(r_t, \hat{r}_{mt}, \hat{\sigma}^2_{mt})]$, even when the losses are possibly contemporaneously and/or serially correlated. The RC and SPA tests are variations of the MDM test, but differ in important ways. First, rather than having as objective to test whether each of the comparison models is significantly better than the benchmark, their main objective is to test whether the best model is significantly better or not. Second, the RC and SPA tests take into account that the data is re-used in the search for a model. The main difference between the SPA and RC tests is that they use different test-statistics, and according to Hansen (2005) the SPA test is more powerful and less sensitive to irrelevant alternatives than the RC test.

The purpose here is to assess the power of the MDM, RC and SPA tests under the alternative, that is, how often the null is rejected given that one or more models are better than the benchmark. The benchmark model that will be used is model 4, the constant variability model, which means that the three other models 1, 2 and 3 are better by construction. Three loss functions $g(\cdot)$ will be studied in the simulations, MSE, MAE and kurtosis. The loss differential d_t at t is defined as $g(r_t, \hat{r}_{4t}, \hat{\sigma}^2_{4t}) - g(r_t, \hat{r}_{mt}, \hat{\sigma}^2_{mt})$. When MSE is the loss function then $d_t = (r^2_t - \hat{r}^2_{4t} - \hat{\sigma}^2_{4t})^2 - (r^2_t - \hat{r}^2_{mt} - \hat{\sigma}^2_{mt})^2$, where $r^2_t - \hat{r}^2_{4t} - \hat{\sigma}^2_{4t}$ is the variability forecast error of model 4 at t, and where $r^2_t - \hat{r}^2_{mt} - \hat{\sigma}^2_{mt}$ is the variability forecast error of model m at t. When MAE is the loss function then $d_t = |r^2_t - \hat{r}^2_{4t} - \hat{\sigma}^2_{4t}| - |r^2_t - \hat{r}^2_{mt} - \hat{\sigma}^2_{mt}|$, and when kurtosis is the loss function, then the loss-differential d_t is equal to $\hat{z}^4_{4t} - \hat{z}^4_{mt}$.

Table 10 contains the simulated rejection probabilities of the nulls of equal or greater loss 1-step ahead for various sample sizes T, using a nominal size of 10%, for the benchmark values $\mathbf{a}_1 = (5^{-1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$ of the simulation DGP. For MDM three tests are made at each sample size, namely m_1 against m_4, m_2 against m_4 and m_3 against m_4. Ideally, since the models have been specified such that m_1 is better than m_2, m_2 is better than m_3 and m_3 is better than m_4, the MDM results should exhibit three properties. First, that the rejection probability of m_1 vs. m_4 is equal to or higher than the rejection probability of m_2 vs. m_4, and that the rejection probability of m_2 vs. m_3 is equal or higher than the rejection probability of m_3 vs. m_4. This property is desirable because multiple comparison tests are often used to choose among models, and so it is desirable that better models are more likely to reject the null. The table suggests that MSE satisfies this property at all sample sizes, although the probabilities of m_2 vs. m_4 and m_3 vs. m_4 are virtually equal at all sample sizes. MAE is close to satisfying the property, since the rejection probability of m_1 vs. m_4 is always higher than the two other rejection probabilities. However, although the rejection probability of m_2 vs. m_4 is similar to the rejection probability of m_3 vs. m_4 at all sample sizes (the biggest difference is 2% points), the latter is always greater or equal. Kurtosis by contrast does not exhibit an increasing rejection probability, since it always produces a rejection probability of zero. This adds to the previous evidence that the standardised residual kurtosis is inappropriate for model comparison, or at least when the mean specification differs. A second property that is desirable for a multiple comparison test is that the rejection probabilities increase with sample size. Both MSE and MAE exhibit this property, but kurtosis does not. A third property that a multiple comparison test should ideally exhibit is sufficiently high power to reject the null in small samples. The table suggests indeed that this is the case with MAE for the benchmark values, since the probability is more than 50% when $T = 50$, and more than 70% when $T = 100$. Unfortunately, additional simulations (not reported in the tables) suggest that these probabilities can be substantially lower for different parameter values. For example, with no mean, that is, $\mathbf{a} = (0, 0.02, 0.1, 0.8, 0.2, 0.1)$, the maximum MAE probabilities are

22% for $T = 25$, 26% for $T = 50$, 30% for $T = 100$, 53% for $T = 500$ and 73% for $T = 1000$. In other words, when the mean information carries little or no explanatory power the MDM test is unlikely to reject the null in small samples. With a mean but no ARCH and no jump by contrast, that is, $\mathbf{a} = (5^{1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$, the maximum MAE probabilities for the five sample sizes are generally higher, namely 19%, 33%, 59%, 100% and 100%.

For the benchmark values $\mathbf{a}_1 = (5^{-1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$ the results for the SC and SPA tests in table 10 can be summarised in five characteristics: (1) the rejection probabilities generally increase when sample size increases, (2) for both RC and SPA the MAE criterion is more powerful than MSE and kurtosis, (3) the power of the RC and SPA tests are relatively similar, since they differ a maximum of 10% points (for MSE when $T = 100$), (4) the RC test is more powerful than SPA when MSE is used as criterion, whereas SPA is more powerful when MAE is used, and (5) both RC and SPA are powerful in small samples, since their rejection probability is about 50% for $T = 50$ and $T = 100$ when MAE is used. Unfortunately, however, additional simulations (not reported in the tables) suggest that the characteristics (1)-(5) are not necessarily reproduced when the parameter values differ from the benchmark values. In particular, with no mean ($\mathbf{a} = (0, 0.02, 0.1, 0.8, 0.2, 0.1)$) the maximum rejection probabilities are 22% , 26% , 30% , 53% and 73% for the five sample sizes, whereas with a mean but no ARCH and no jump ($\mathbf{a} = (5^{1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$) the SPA probabilities using MAE are 19%, 33%, 59%, 100% and 100%. This suggests that the power to reject the null can be very low in small samples, in particular when there is little or no explanatory information in the mean, and/or if there is little or no explanatory information in the variance. A possibly even more serious shortcoming suggested by the additional simulations is that they do not seem to provide any clear guidance as to whether MSE or MAE is preferable, since their comparative power depend greatly on the parameter values of the DGP. Similarly, the additional simulations to not provide clear guidance as to whether RC or SPA is preferable nor under which circumstances. Possibly a more comprehensive and detailed simulation study could shed further light on these questions.

3.5 Mincer-Zarnowitz regressions

The loss functions provide information about the ranking between models, whereas the multiple comparison tests provide information about whether any model or group of models is significantly better than the benchmark model(s). However, neither the loss functions nor the tests provide information about the degree of forecast bias. Several tests associated with Mincer-Zarnowitz regressions provide simple ways of obtaining such information. Mincer-Zarnowitz regressions of variability r_t^2 on variability forecasts take the form

$$r_t^2 = a + b(\hat{r}_{mt}^2 + \hat{\sigma}_{mt}^2) + u_{mt}, \qquad t = 1, \dots, T \tag{14}$$

where $\hat{r}_{mt}^2 + \hat{\sigma}_{mt}^2$ is the return variability forecast of model m. Ideally a and b should be equal to 0 and 1, respectively, because then the forecasts are deemed "unbiased" in the sense that they do not tend to over- nor underpredict. Table 9 contains the simulated rejection probabilities of four different null hypotheses associated with Mincer-Zarnowitz regressions, using a nominal level of 10%. It should be noted that for model 4 it is not possible to undertake tests 1 and 4, since the variability forecasts of model 4 are constant (including a constant in addition to variability as regressor produces co-linearity). In test 1 the null is $a = 0$ and should not be rejected for model 1, whereas it should be rejected for models 2 and 3. The rejection probabilities are usefully close to the nominal level of 10% for model 1, since they range from 22% when $T = 25$ to 11% when T is equal to 1000. Also, for $T = 50$ or higher the rejection probability falls as the sample size increases. For model 2 and 3 the rejection probability increases—as desired—with sample size, but unfortunately the probabilities are somewhat low in small samples since they vary from 27% when $T = 25$ to 48% when $T = 100$ for model 2, and from 32% when $T = 25$ to 38% when $T = 100$ for model 3. This suggests test 1 might not be very informative in practice in small samples.

Tests 2 and 3 do not exhibit desirable properties. In test 2 the null is $b = 0$ and it would be desirable that the null is rejected for model 1, and that model 1 exhibits the highest rejection probability for each sample size. Compared with model 2 and 3 this is indeed the case, but not compared with model 4. Similarly, in test 3 it would be desirable that the null of $b = 1$ is rejected for models 2, 3 and 4 but not for model 1, and that the probabilities increase with sample size T for the former models and decreases with T for the latter. Unfortunately, this is not the case for model 2 where the rejection probabilities decrease with sample size, and where rejection probabilities are lower than for model 1 (except when $T = 1000$).

In test 4 the null is the joint hypothesis that $a = 0$ and $b = 1$. As in test 1 the rejection probabilities decrease with sample size for model 1, and in large samples the rejection probability is close to the nominal level as they range from 21% for $T = 100$ to 11% for $T = 1000$. In small samples, however, the test is notably oversized since the probabilities are 36% for $T = 25$ and 28% for $T = 50$. A property of test 4 which is in line with the results for test 1, is that the rejection probabilities of models 2 and 3 generally increase—as desired—with sample size. The qualifier "generally" refers to the property that, for model 3, probabilities first fall until $T = 100$ and then increases. Overall, then, the results suggest that tests 1 and 4 can be useful in econometric practice, but the degree of usefulness depends on sample size. For small samples (T between 25 and 100) the tests may not be very informative. Also, Mincer-Zarnowitz regressions on constant variability predictions are not very useful, since tests 1 and 4 are not applicable to constant models of variability.

3.6 A simple framework

The simulations suggest the following simple three step framework for financial return variability point forecast comparison. First, use MAE or MSE of variability forecasts to rank the models, since both MAE and MSE exhibit the property that the probability of correctly ranking each model—regardless of the others' ranking—is always the highest among the rank probabilities. Overall, these properties are retained also for parameter values that differ from the benchmark DGP. Whether MAE or MSE is more appropriate depends on sample size. Generally, MAE is more likely to provide the correct ranking in small samples up to about 100 observations, whereas MSE is more likely to provide the correct ranking in samples larger than 100 observations. With MSE the probability of obtaining the correct ranking increases with sample size. With MAE the probability of obtaining the correct ranking increases with sample size until $T = 500$, but then decreases as the sample size increases further. Closer inspection of the simulation results revealed that the source of this unexpected behaviour is the constant model, whose ranking probability decreases when the sample size becomes very large. The second step of the framework consists of comparing the models against a benchmark using the MDM test and the RC and/or SPA tests. For the MDM test the MAE is generally more powerful than MSE and kurtosis, and this property remains for parameter values that differ from the benchmark DGP. The properties of the RC and SPA tests by contrast depend on the parameter values of the simulation DGP. Finally, the third step of the framework consists of running Mincer-Zarnowitz regressions, focusing on the R^2 of the regressions and on the joint hypothesis test $a = 0, b = 1$. The test provides information about the degree of forecast bias and in the simulations it exhibited two desirable properties. Namely that the rejection probabilities tend toward the nominal size as the sample size increases when the null is true, and that the rejection probability generally increases with sample size when the null is not true.[4] The R^2 provides additional information on bias and how it can possibly be corrected. For example, if a model ranks badly according to MSE and MAE but produces a high R^2, then this suggests that the model's forecasts can be considerably improved upon simply through a linear transformation. Indeed, an example of this is RV in the *ex post* comparison in the next section.

4 An empirical illustration

The purpose of this section is to illustrate the use in practice of the simple framework outlined at the end of the previous section. The illustration will be on weekly (close, Friday-to-Friday) Norwegian exchange rate (NOK/EUR) data from 7 October 2005 to 4 January 2008, a total of 118 weekly observations. The reason behind

[4] "Generally" because the simulations suggests model 3 is an exception. For model 3 the rejection probability first decreases until $T = 100$ before increases again.

this data choice is that they are very suited to illustrate the methodological and practical issues that can arise in the forecast evaluation of explanatory models of financial variability. The Norwegian krone is a minor currency in terms of volume in the currency markets, and so market microstructure issues are likely to be more pronounced than for, say, the EUR/USD exchange rate. Also, Norges Bank (The Central Bank of Norway) makes weekly order flow data of the Norwegian krone (NOK) freely available on their website.[5] In order to undertake a true out-of-sample forecast evaluation the sample is divided in two at 19 January 2007. The 68 observations up to and including this date constitute the estimation and model design sample, whereas the 50 observations after this date constitute the forecast evaluation sample. No re-estimation of any model is undertaken after 19 January 2007, so the experiment is a true out-of-sample exercise. Both *ex post* and *ex ante* evaluations are undertaken, but for simplicity only for 1-step forecasts. The objective of an *ex post* evaluation is to shed light on the accuracy in conditional forecasting and counterfactual analysis situations. In other words, how well an explanatory model forecasts given that the values of the conditioning variables are correct. If correctly predicting the values of the conditioning variables does not improve upon forecast accuracy beyond that of the non-explanatory models, then this suggests the explanatory model does not constitute an improvement in conditional forecasting and counterfactual analysis compared with the non-explanatory models. The objective of an *ex ante* evaluation is to shed light on the accuracy of explanatory models when the values of the conditioning variables are uncertain. One cannot necessarily expect the explanatory model to forecast better than the non-explanatory models in this case, but ideally the explanatory model should forecast at least *as well* as the non-explanatory models.

The section is divided in two. The first subsection presents the models to be compared, whereas the second subsection contains the out-of-sample forecast evaluation.

4.1 The models

Four models of exchange rate return variability are compared. The first model is an explanatory model of exchange rate return and is referred to as ECON. The model is explanatory in the sense that it contains several explanatory variables, including currency order flow, and Norwegian and euro-area money market interest rates. The model is given by:

[5] The data are collected daily since 2 October 2005, but Norges Bank only make weekly aggregates publicly available via their statistics webpages. More precisely, the data can be downloaded via the url `http://www.norges-bank.no/Pages/ReportRoot____60389.aspx` and are described in more detail in Meyer and Skjelvik (2006).

$$
\begin{aligned}
\Delta s_t &= \underset{[0.00]}{0.09}\Delta x_t - \underset{[0.02]}{1.48}(\Delta ir_t^{no} - \Delta ir_t^{emu}) - \underset{[0.00]}{17.11}ECM_{t-1} + \hat{e}_{1t}, \\
ECM_t &= s_t - 1.99 + 4.51 ir_t^{no} - 8.02 ir_t^{emu} \\
\hat{e}_{1t} &= \hat{\sigma}_{1t}\hat{z}_{1t}, \quad \hat{\sigma}_{1t} = 0.57, \quad \hat{z}_{1t} \sim IIN(0,1)
\end{aligned}
$$

$$
R^2\ 0.42 \quad AR_1\ \underset{[0.31]}{1.02} \quad ARCH_1\ \underset{[0.90]}{0.02} \quad JB\ \underset{[0.99]}{0.03} \quad T = 68
$$

The variable Δs_t is the Norwegian krone against the euro (NOK/EUR) log-return in percentages from the end of Friday in week $t-1$ to the end of Friday in week t, which means positive values implies a depreciation of the Norwegian krone. Δx_t is a measure of worldwide forward order flow involving the Norwegian krone (positive values means there is net demand for foreign currency) in billions of Norwegian kroner, Δir_t^{no} is the change in the Norwegian 1-week money market yield in percentage points and Δir_t^{emu} is the change in the euro-area 1-week money market yield in percentage points.[6] The term ECM_t is the estimated disequilibrium implied by an OLS estimated cointegration relation between s_t, ir_t^{no} and ir_t^{emu}, where s_t is equal to log(NOK/EUR). The explanatory power in terms of R^2 is 0.42, which is high in an exchange rate context, and the errors are homoscedastic and normal according to standard tests and common significance levels.[7] Several conditional variance specifications that included GARCH terms and explanatory variables—including volume variables—were tried, all resulting in either numerical problems or insignificant parameter estimates. So even though the conditional variance might not be homoscedastic, this is nevertheless the practical option that suggests itself to the modeller according to standard tests and modelling strategies. ECON's *ex post* forecast $\hat{r}_{1t}$ of the conditional mean is given by $0.09\Delta x_t - 1.48(\Delta ir_t^{no} - \Delta ir_t^{emu}) - 17.11ECM_{t-1}$, and the *ex post* forecast of conditional variability is given by $\hat{r}_{1t}^2 + \hat{\sigma}_{1t}^2$. In an *ex ante* situation the contemporaneous values of the variables Δx_t, Δir_t^{no} and Δir_t^{emu} would have to be forecast. For simplicity, the *ex ante* forecast of the squared conditional mean forecast $\hat{r}_{1t}^2$ is specified as $(0.09)^2\overline{x} + (1.48)^2\overline{ir} + (17.11ECM_{t-1})^2$, where $\overline{x}$ and $\overline{ir}$ are the sample variances of Δx_t and $(\Delta ir_t^{no} - \Delta ir_t^{emu})$, respectively, in the estimation and model design sample.

The second model that will be evaluated is realised volatility (RV), that is, the sum of squared intra-weekly equidistant returns. Under certain assumptions,

[6] The rawdata of s_t, ir_t^{no} and ir_t^{emu} are the daily series ew:nor19101, ew:14307 and ew:emu14813 from EcoWin. The source of and further reading on the order flow data is contained in footnote 5.

[7] AR_1 and $ARCH_1$ are the Ljung and Box (1979) test statistic for first order serial correlation in the residuals and squared residuals, respectively, and JB is the Jarque and Bera (1980) test statistic for non-normality in the residuals. Values in square brackets are the p-values associated with the tests.

including no measurement error and market microstructure noise, it can be showed that RV provides a consistent estimate of quadratic variation (QV)—a continuous time analogue of discrete time volatility—when the time increment goes to zero (recall the discussion in section 2). The assumptions of no measurement error and no market microstructure noise are unlikely to hold—in particular in the Norwegian case, and numerous modifications and extensions to RV have been proposed, see Aït-Sahalia (2006) for an overview. For simplicity, however, only RV is included here. Our weekly RV series is made up of 30 minute squared returns using end-of-interval mid-point quotes from Olsen Financial Technologies (OFT). The RV model is given by:

$$\begin{aligned}
\Delta s_t &= \hat{e}_{2t}, \quad \hat{e}_{2t} = \hat{\sigma}_{2t}\hat{z}_{2t}, \quad \hat{z}_{2t} \sim IIN(0,1) \\
\hat{\sigma}^2_{2t} &= \sum_{n(t)=1}^{N(t)} (\Delta s_{n(t)})^2
\end{aligned}$$

$$R^2 \; 0.00 \quad AR_1 \underset{[0.80]}{0.06} \quad ARCH_1 \underset{[0.38]}{0.78} \quad JB \underset{[0.68]}{0.77} \quad T = 68$$

The term $\hat{\sigma}^2_{2t}$ is RV at t and the diagnostic tests AR_1, $ARCH_1$ and JB are of the standardised residual $\hat{z}_{2t} = \Delta s_t / \hat{\sigma}_{2t}$. The *ex post* variability forecast is given by $\hat{\sigma}^2_{2t}$, that is, RV at t, whereas the *ex ante* forecast is given by the fitted values of an AR(1) model of RV.[8]

The third model is a plain exponential GARCH(1,1) model. The model is "plain" in the sense that the conditional mean is set to zero, and the model is exponential in the sense that the conditional variance has an exponential specification. The main motivation for the exponential specification instead of a GARCH(1,1) is that the latter produces negative fitted values of conditional variance. The EGARCH(1,1) is widely used and has been extensively studied in the academic literature since it was put forward by Nelson (1991). For simplicity in estimation the EGARCH(1,1) model used here differs slightly from Nelson's original model, and specifically it is given by:

$$\begin{aligned}
\Delta s_t &= \hat{e}_{3t}, \quad \hat{e}_{3t} = \hat{\sigma}_{3t}\hat{z}_{3t}, \quad \hat{z}_{3t} \sim IIN(0,1) \\
\log \hat{\sigma}^2_{3t} &= -\underset{[0.00]}{1.19} + \underset{[0.48]}{0.25} \left| \frac{\hat{e}_{3t-1}}{\hat{\sigma}_{3t-1}} \right| - \underset{[0.37]}{0.50} \log \hat{\sigma}^2_{3t-1}
\end{aligned}$$

$$R^2 \; 0.00 \quad AR_1 \underset{[0.73]}{0.12} \quad ARCH_1 \underset{[0.80]}{0.06} \quad JB \underset{[0.55]}{1.19} \quad T = 68$$

[8] Only one lag is included because further lags are insignificant at 10%. The in-sample R^2 of the fitted model is 16%, and the standardised residuals are non-normal white noise according to standard diagnostic tests.

The diagnostic tests are of the standardised residuals, and both the *ex post* and *ex ante* forecasts of variability are given by $\hat{\sigma}^2_{3t}$.[9]

The fourth and final model that is included in the comparison is a constant variance model, and here it takes the form of the simplest version of the sample variance (variation about zero, division over T):

$$\begin{aligned}\hat{\sigma}^2_{4t} &= \frac{1}{T}\sum_{t=1}^{T}(\Delta s_t)^2 \\ & R^2\ 0.00 \quad AR_1\ \underset{[0.99]}{0.00} \quad ARCH_1\ \underset{[0.86]}{0.03} \quad JB\ \underset{[0.49]}{1.44} \quad T = 68\end{aligned}$$

The diagnostic tests are of the standardised residual $\hat{z}_{4t} = r_t/\hat{\sigma}_{4t}$, and both the *ex post* and *ex ante* forecasts are given by $\hat{\sigma}^2_{t4}$.

4.2 Variability forecast comparison

Explanatory models can provide conditional forecasts, say, the impact on variability of a change in the interest rate, and counterfactual analysis, say, what would the profit had been if a derivative had been priced conditional on a change in the interest rate rather than not. The objective of an *ex post* comparison is to evaluate the forecast accuracy of explanatory models in such situations, which amounts to the assumption that the values of the conditioning variables are correct. If explanatory models do not fare better than the "non-explanatory" models when the conditioning information is correct, then the explanatory models do not provide insight beyond the non-explanatory models for such purposes. Table 11 contains the *ex post* 1-step out-of-sample forecast evaluation results of the four models. According to the MAE criterion ECON is the best forecaster of variability, the constant variability model is second, the EGARCH(1,1) comes third and RV is last. The p-value of the SPA test is 13%, which suggests that there is no model that is significantly better than the benchmark MAE at common significance levels. However, one should have in mind that the supporting simulations suggested that the power of the SPA test can be very low in small samples—even when the mean information carries a reasonably high explanatory power. The MDM test suggests stronger insignificance, since the lowest p-value produced by the three models that are tested against the benchmark is 31%. But also for the MDM test should one keep in mind that the power can be very low. That RV produces the worst *ex post* forecast of variability according to MAE is possibly surprising, but an explanation is suggested by the relative high R^2 of 20% in the Mincer-Zarnowitz regression. This is second highest after ECON with 0.48%, which suggests that the RV forecast is biased and can be improved upon in a

[9] Estimation is by quasi maximum likelihood (QML) in EViews 6 using the Marquardt algorithm and no backcasting.

straightforward manner.[10] The bias of the RV forecast nevertheless underlines that market microstructure effects can seriously affect the precision of high-frequency estimates that are based on continuous time theory. The R^2 of the GARCH(1,1) model's forecasts are as expected very low, 1%, whereas the R^2 of the constant model's forecast by construction is equal to zero. The joint Wald test of $a = 0, b = 1$ is not rejected for ECON, whereas it is for RV and EGARCH(1,1). Moreover, the residuals of RV are serially correlated, which supports the previous evidence of it being biased.[11] All in all, then, the results are indeed in favour of ECON for conditional forecasting and counterfactual analysis purposes.

Ideally an explanatory model should not only be useful for conditional forecasting and counterfactual analysis in situations where the assumption that the values of the conditioning variables are correct is appropriate. Explanatory models should also be useful for forecasting when the values of the conditioning variables are uncertain. One cannot expect that explanatory models fare better than non-explanatory models in such cases. However, it is desirable that they fare at least *as good* as non-explanatory models. Table 12 contains the *ex ante* 1-step out-of-sample forecast evaluation results of the four models. It should be noted that for the EGARCH(1,1) and constant models the *ex post* and *ex ante* forecasts are the same. According to MAE the constant model is best, ECON is second, the EGARCH(1,1) is third whereas RV is last. Unsurprisingly, therefore, the SPA test suggests that none of the comparison models have a significantly smaller MAE than the constant model (and similarly for the MDM test). The R^2s are low and equal to 1% for ECON, RV and EGARCH(1,1), which is common in *ex ante* forecasting of variability. Turning to the joint test of $a = 0, b = 1$ in the Mincer-Zarnowitz regression, that is, the bias test, then for all of the three models in which it can be undertaken is it rejected. Overall, then, although the results do not point to a clear winner, the evidence is in favour of the constant variability model. Put differently, the *ex post* results suggest ECON should be used in scenario analysis, say, conditional forecasting, whereas the *ex ante* results suggest the constant model should be used in *ex ante* forecasting.

5 Conclusions

Evaluating explanatory models of financial inter-period return variability against high-frequency intra-period estimates based on continuous time theory raises several methodological and practical issues. Together these methodological and practical issues suggest that an alternative framework is needed when comparing explanatory models' forecasts of financial return variability. This study has contributed to this area in two ways. First, the finite sample properties of operational and practical procedures for the evaluation of explanatory discrete time models of financial return

[10] Also according to MSE (not reported) does the RV come second after ECON.

[11] The p-value of the Wald test for RV does not change when Newey and West (1987) estimates are used in order to account for serially correlated residuals.

variability has been studied, where return variability is defined as squared return. Second, with basis in the simulation results a simple framework for the evaluation of explanatory models of return variability has been proposed and illustrated.

The simple framework contains three steps. First, compute the MAE or MSE variability forecast errors. Whether the MAE or MSE is more appropriate depends on sample size. The simulations suggest that MAE is more appropriate when the sample size is lower than about 100 observations, whereas MSE is more appropriate when the sample size is higher. The second step of the framework consists in testing for significantly superior forecasts using the MDM and SPA/RC tests. The tests exhibit relatively high power for values in the simulation DGP for the benchmark values—even in small samples. However, when the models differ less and/or when the mean and/or variance specification account for little of the conditional variation, then the power can be very low—in particular in small samples. This should be kept in mind when interpreting the output of the tests, in particular when they suggest insignificant superior forecast precision. The third and final step consists in testing for forecast bias by means of a Mincer-Zarnowits regression of actual value on a constant and the forecast, paying particular attention to the joint restriction test of the constant being equal to zero and the slope coefficient being equal to one.

The results of this study can be investigated further and complemented in many ways, but here only two suggestions are given. First, although the benchmark values of the simulation DGP were carefully selected with a view to the empirical properties financial returns actually exhibit, further study is needed. In particular, further investigation is needed in order to better understand how the loss functions and statistical tests behave when when the standardised residual is more fat-tailed than the Gaussian distribution. Second, although the MDM, RC, SPA and Mincer-Zarnowitz tests performed reasonably well in small samples for the benchmark values, the power decreases substantially when the explanatory information in either the mean or variance specifications tends to zero. A test with more power in small samples is desirable, and the forecast evaluation literature contains a large number of potential candidates that can possibly be evaluated.

References

Aït-Sahalia, Y. (2006). Estimating Continuous-Time Models with Discretely Sampled Data. Invited lecture at the 2005 World Congress of the Econometric Society. To appear as chapter 9 in volume 3 of the conference proceedings, and currently available via `http://cemmap.ifs.org.uk/`.

Aït-Sahalia, Y. and P. A. Mykland (2003). The Effects of Random and Discrete Sampling When Estimating Continuous-Time Diffusions. *Econometrica 71*, 483–549.

Aït-Sahalia, Y., P. A. Mykland, and L. Zhang (2005). How Often to Sample a

Continuous-Time Process in the Presence of Market Microstructure Noise. *Econometrica 71*, 483–549.

Andersen, T., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold (2006). Volatility and correlation forecasting. In G. Elliott, C. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting, Volume 1*. Amsterdam: North Holland.

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

Andersen, T. G., T. Bollerslev, F. S. Diebold, and P. Labys (2001). The Distribution of Realized Exchange Rate Volatility. *Journal of the American Statistical Association 96*, 42–55. Correction published in 2003, volume 98, page 501.

Andersen, T. G., T. Bollerslev, F. S. Diebold, and P. Labys (2003). Modeling and Forecasting Realized Volatility. *Econometrica 72*, 579–625.

Andersen, T. G., T. Bollerslev, and S. Lange (1999). Forecasting Financial Market Volatility: Sample Frequency vis-à-vis Forecast Horizon. *Journal of Empirical Finance 6*, 457–477.

Andersen, T. G., T. Bollerslev, and N. Meddahi (2005). Correcting the Errors: Volatility Forecast Evaluation Using High-Frequency Data and Realized Volatilities. *Econometrica 73*, 279–296.

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

Barndorff-Nielsen, O. E. and N. Shephard (2002b). Estimating Quadratic Variation Using Realized Variance. *Journal of Applied Econometrics 17*, 457–477.

Blume, M. E., A. C. Mackinlay, and B. Terker (1989). Order Imbalances and Stock Price Movements on October 19 and 20, 1987. *The Journal of Finance 44*, 827–848.

Campos, J., N. R. Ericsson, and D. F. Hendry (2005). General-to-Specific Modeling: An Overview and Selected Bibliography. In J. Campos, D. F. Hendry, and N. R. Ericsson (Eds.), *General-to-Specific Modeling, Volume 1*. Cheltenham: Edward Elgar Publishing.

Chordia, T., R. Roll, and A. Subrahmanyam (2002). Order imbalance, liquidity, and market returns. *Journal of Financial Economics 65*, 111–130.

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

Dunne, P., H. Hau, and M. Moore (2005). International Order Flows: Explaining Equity and Exchange Rate Returns. Presented at the Norges Bank Conference on Microstructure in Oslo, September 2005.

Engle, R. F. and A. J. Patton (2004). Impacts of trades in an error-correction model of quote prices. *Journal of Financial Markets 7*, 1–25.

Escribano, A. and R. Pascual (2006). Asymmetries in bid and ask responses to innovations in the trading process. *Empirical Economics 30*, 913–946.

Evans, M. D. and R. K. Lyons (2002). Order flow and exchange rate dynamics. *Journal of Political Economy 110*, 170–180.

Florens, J.-P., M. Mouchart, and J.-F. Richard (1990). *Elements of Bayesian Statistics.* New York: Marcel Dekker.

Gilbert, C. L. (1990). Professor Hendry's Econometric Methodology. In C. W. Granger (Ed.), *Modelling Economic Series.* Oxford: Oxford University Press. Earlier publised in Oxford Bulletin of Economics and Statistics 48 (1986), pp. 283-307.

Hansen, P. R. (2005). A Test for Superior Predictive Ability. *Journal of Business and Economic Statistics 23*, 365–380.

Hansen, P. R. and A. Lunde (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics 20*, 873–889.

Hansen, P. R. and A. Lunde (2006). Consistent ranking of volatility models. *Journal of Econometrics 131*, 97–121.

Hansen, P. R. and A. Lunde (2007). MULCOM 1.00. Econometric Toolkit for Multiple Comparisons. `http://www.asb.dk/~alunde/mulcom/mulcom.htm`.

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting 23*, 801–824.

Hasbrouck, J. (1991). Measuring the Information Content of Stock Trades. *The Journal of Finance 46*, 179–207.

Hendry, D. F. (1995). *Dynamic Econometrics.* Oxford: Oxford University Press.

Hendry, D. F. and J.-F. Richard (1982). On the Formulation of Empirical Models in Dynamic Econometrics. *Journal of Econometrics 20*, 3–33.

Jarque, C. and A. Bera (1980). Efficient Tests for Normality, Homoskedasticity, and Serial Independence of Regression Residuals. *Economics Letters 6*, 255–259.

Lee, C. L. and M. J. Ready (1991). Inferring Trade Direction from Intraday Data. *The Journal of Finance 46*, 733–746.

Ljung, G. and G. Box (1979). On a Measure of Lack of Fit in Time Series Models. *Biometrika 66*, 265–270.

Meddahi, N. (2002). A Theoretical Comparison Between Integrated and Realized Volatility. *Journal of Applied Econometrics 17*, 479–508.

Meyer, E. and J. Skjelvik (2006). Statistics on foreign exchange transactions — new insight into foreign exchange markets. *Norges Bank Economic Bulletin* (2/06), 80–88. Available as `http://www.norges-bank.no/upload/import/english/publications/economic_bulletin/2006-02/meyer.pdf`.

Mincer, J. and V. Zarnowitz (1969). The Evaluation of Economic Forecasts. In J. Zarnowitz (Ed.), *Economic Forecasts and Expectations*. New York: National Bureau of Economic Research.

Mizon, G. (1995). Progressive Modeling of Macroeconomic Time Series: The LSE Methodology. In K. D. Hoover (Ed.), *Macroeconometrics. Developments, Tensions and Prospects*. Kluwer Academic Publishers.

Moberg, J. and G. Sucarrat (2007). Stock Market Return, Order Flow and Financial Markets Linkages. Work in progress, current version available as `http://www.eco.uc3m.es/sucarrat/research/stocknflows.pdf`.

Nelson, D. B. (1991). Conditional Heteroscedasticity in Asset Returns: A New Approach. *Econometrica 51*, 485–505.

Newey, W. and K. West (1987). A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica 55*, 703–708.

Patton, A. J. (2007). Volatility Forecast Evaluation and Comparison Using Imperfect Volatility Proxies. Available as `http://www.economics.ox.ac.uk/members/andrew.patton/patton_robust_aug07.pdf`.

Sucarrat, G. (2007). Econometric Reduction Theory and Philosophy: The First Stage in Hendry's Reduction Theory Revisited. Available as `http://www.eco.uc3m.es/sucarrat/research/hendry.pdf`.

Taylor, S. J. and X. Xu (1997). The incremental information in one million foreign exchange quotations. *Journal of Empirical Finance 4*, 317–340.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity. *Econometrica 48*, 817–838.

White, H. (2000). A Reality Check for Data Snooping. *Econometrica 68*, 1097–1126.

Zhou, B. (1996). High-Frequency Data and Volatility in Foreign-Exchange Rates. *Journal of Business and Economic Statistics 14*, 45–52.

Table 1: Descriptive statistics of interdaily (close, weekends excluded) exchange rate returns in percent from 26 September 2005 to 4 January 2008 ($T = 594$)

T	MSE	MAE	Kurtosis	R^2
25	0.28	0.44	0.01	0.33
50	0.39	0.49	0.01	0.47
100	0.52	0.57	0.01	0.65
500	0.85	0.62	0.00	0.96
1000	0.96	0.61	0.00	0.99

Note: $S.E.$ is the standard error of returns, $Kurtosis$ is the sample estimate of kurtosis, JB is the Jarque and Bera (1980) test for non-normality, and $AR(1)$ and $ARCH(1)$ are Ljung and Box (1979) tests for first order serial correlation in returns and squared returns, respectively. Values in square parentheses are the p-values associated with the tests.

Table 2: Descriptive statistics of weekly (close, Friday-to-Friday) exchange rate returns in percent from 30 September 2005 to 4 January 2008 ($T = 118$)

	USD/EUR	YEN/EUR	GBP/EUR	NOK/EUR
$S.E.$	1.021	1.170	0.740	0.780
$Kurtosis$	2.666	7.047	3.081	4.033
JB	0.550	118.490	0.166	10.845
	[0.76]	[0.00]	[0.92]	[0.00]
$AR(1)$	2.045	5.799	0.052	0.029
	[0.15]	[0.02]	[0.82]	[0.86]
$ARCH(1)$	0.098	3.519	0.920	1.158
	[0.75]	[0.06]	[0.34]	[0.28]

Note: $S.E.$ is the standard error of returns, $Kurtosis$ is the sample estimate of kurtosis, JB is the Jarque and Bera (1980) test for non-normality, and $AR(1)$ and $ARCH(1)$ are Ljung and Box (1979) tests for first order serial correlation in returns and squared returns, respectively. Values in square parentheses are the p-values associated with the tests.

Table 3: Descriptive statistics of simulated returns for different parameter values $\mathbf{a}_l, l = 1, \ldots, 5$

	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_4$	$\mathbf{a}_5$
$S.E.$	0.769	0.629	0.631	0.443	0.446
$Kurtosis$	3.088	2.967	2.942	3.033	2.943
JB	2.521	1.940	1.832	2.340	1.782
R^2	0.347	0.507	0.500	0.000	0.000
R^2 variability	0.019	0.027	0.027	0.011	0.000

Note: Simulations are in EViews 6 with 10 000 replications, each with $T = 100$ and a prior burn-in sample of 100 observations in order to avoid initial value issues. The parameter values of the simulations are $\mathbf{a}_1 = (5^{-1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$, $\mathbf{a}_2 = (5^{-1/2}, 0.02, 0.1, 0.8, 0, 0)$, $\mathbf{a}_3 = (5^{-1/2}, 0.2, 0, 0, 0, 0)$, $\mathbf{a}_4 = (0, 0.02, 0.1, 0.8, 0, 0)$ and $\mathbf{a}_5 = (0, 0.2, 0, 0, 0, 0)$. $S.E.$ is the average standard error of the simulated returns, $Kurtosis$ is the average sample kurtosis, JB is the average Jarque and Bera (1980) test-statistic for non-normality, R^2 is the average R^2 of the OLS regression $r_{lt} = \hat{\gamma}_0 + \hat{\gamma}_1 x_t + \hat{e}_{lt}$, and R^2 *variability* is the average R^2 of the OLS regression $r_{lt}^2 = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{r}_{lt}^2 + \hat{e}_{lt}$, where $\hat{r}_{lt}^2$ is conditional variability of returns for $\mathbf{a}_l$.

Table 4: Probabilities of obtaining a correct ranking of models 1 to 4 using MSE, MAE, Kurtosis and the R^2 of Mincer-Zarnowitz regressions when $\mathbf{a}$ is equal to the benchmark values $(5^{-1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$

T	MSE	MAE	Kurtosis	R^2
25	0.28	0.44	0.01	0.33
50	0.39	0.49	0.01	0.47
100	0.52	0.57	0.01	0.65
500	0.85	0.62	0.00	0.96
1000	0.96	0.61	0.00	0.99

Note: Simulations are in EViews 6 and R 2.6.1 with 1000 replications, each with a prior burn-in sample of 100 observations in order to avoid initial value issues.

Table 5: Ranking probabilities for model 1 using MSE, MAE, Kurtosis and the R^2 of Mincer-Zarnowitz regressions when **a** is equal to the benchmark values $(5^{-1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$

T	Rank	MSE	MAE	Kurtosis	R^2
25	1st.	0.67	0.70	0.11	0.68
	2nd.	0.10	0.10	0.40	0.17
	3rd.	0.16	0.18	0.14	0.16
	4th.	0.08	0.03	0.35	0.00
50	1st.	0.78	0.80	0.05	0.86
	2nd.	0.07	0.08	0.48	0.07
	3rd.	0.12	0.12	0.16	0.08
	4th.	0.04	0.01	0.30	0.00
100	1st.	0.88	0.88	0.05	0.95
	2nd.	0.05	0.06	0.51	0.04
	3rd.	0.07	0.06	0.16	0.02
	4th.	0.01	0.00	0.28	0.00
500	1st.	1.00	1.00	0.00	1.00
	2nd.	0.00	0.00	0.70	0.00
	3rd.	0.00	0.00	0.14	0.00
	4th.	0.00	0.00	0.16	0.00
1000	1st.	1.00	1.00	0.00	1.00
	2nd.	0.00	0.00	0.78	0.00
	3rd.	0.00	0.00	0.11	0.00
	4th.	0.00	0.00	0.10	0.00

Note: Simulations are in EViews 6 and R 2.6.1 with 1000 replications, each with a prior burn-in sample of 100 observations in order to avoid initial value issues.

Table 6: Ranking probabilities for model 2 using MSE, MAE, Kurtosis and the R^2 of Mincer-Zarnowitz regressions when $\mathbf{a}$ is equal to the benchmark values $(5^{-1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$

T	Rank	MSE	MAE	Kurtosis	R^2
25	1st.	0.15	0.12	0.01	0.14
	2nd.	0.42	0.56	0.21	0.42
	3rd.	0.34	0.29	0.47	0.44
	4th.	0.09	0.04	0.31	0.00
50	1st.	0.12	0.10	0.02	0.08
	2nd.	0.50	0.57	0.21	0.52
	3rd.	0.32	0.30	0.43	0.40
	4th.	0.06	0.03	0.34	0.00
100	1st.	0.07	0.06	0.01	0.03
	2nd.	0.58	0.62	0.20	0.66
	3rd.	0.32	0.31	0.45	0.31
	4th.	0.03	0.01	0.34	0.00
500	1st.	0.00	0.10	0.00	0.00
	2nd.	0.85	0.66	0.14	0.96
	3rd.	0.14	0.24	0.48	0.04
	4th.	0.00	0.00	0.38	0.00
1000	1st.	0.00	0.20	0.00	0.00
	2nd.	0.96	0.66	0.10	0.99
	3rd.	0.04	0.15	0.54	0.01
	4th.	0.00	0.00	0.36	0.00

Note: Simulations are in EViews 6 and R 2.6.1 with 1000 replications, each with a prior burn-in sample of 100 observations in order to avoid initial value issues.

Table 7: Ranking probabilities for model 3 using MSE, MAE, Kurtosis and the R^2 of Mincer-Zarnowitz regressions when $\mathbf{a}$ is equal to the benchmark values $(5^{-1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$

T	Rank	MSE	MAE	Kurtosis	R^2
25	1st.	0.12	0.17	0.00	0.18
	2nd.	0.32	0.30	0.30	0.42
	3rd.	0.42	0.51	0.37	0.40
	4th.	0.15	0.02	0.33	0.00
50	1st.	0.07	0.10	0.00	0.06
	2nd.	0.32	0.34	0.26	0.42
	3rd.	0.48	0.55	0.39	0.52
	4th.	0.13	0.02	0.35	0.00
100	1st.	0.04	0.09	0.00	0.02
	2nd.	0.32	0.32	0.25	0.30
	3rd.	0.57	0.59	0.38	0.68
	4th.	0.07	0.00	0.37	0.00
500	1st.	0.00	0.04	0.00	0.00
	2nd.	0.14	0.34	0.16	0.04
	3rd.	0.85	0.63	0.38	0.96
	4th.	0.00	0.00	0.46	0.00
1000	1st.	0.00	0.05	0.00	0.00
	2nd.	0.04	0.34	0.12	0.01
	3rd.	0.96	0.62	0.35	0.99
	4th.	0.00	0.00	0.54	0.00

Note: Simulations are in EViews 6 and R 2.6.1 with 1000 replications, each with a prior burn-in sample of 100 observations in order to avoid initial value issues.

Table 8: Ranking probabilities for model 4 using MSE, MAE, Kurtosis and the R^2 of Mincer-Zarnowitz regressions when **a** is equal to the benchmark values $(5^{-1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$

T	Rank	MSE	MAE	Kurtosis	R^2
25	1st.	0.07	0.02	0.88	0.00
	2nd.	0.16	0.04	0.09	0.00
	3rd.	0.09	0.04	0.02	0.00
	4th.	0.69	0.91	0.01	1.00
50	1st.	0.03	0.01	0.93	0.00
	2nd.	0.12	0.02	0.05	0.00
	3rd.	0.08	0.04	0.02	0.00
	4th.	0.77	0.93	0.01	1.00
100	1st.	0.01	0.00	0.97	0.00
	2nd.	0.05	0.00	0.02	0.00
	3rd.	0.05	0.03	0.01	0.00
	4th.	0.89	0.97	0.01	1.00
500	1st.	0.00	0.00	0.99	0.00
	2nd.	0.00	0.00	0.00	0.00
	3rd.	0.00	0.10	0.00	0.00
	4th.	1.00	0.90	0.00	1.00
1000	1st.	0.00	0.00	1.00	0.00
	2nd.	0.00	0.00	0.00	0.00
	3rd.	0.00	0.16	0.00	0.00
	4th.	1.00	0.84	0.00	1.00

Note: Simulations are in EViews 6 and R 2.6.1 with 1000 replications, each with a prior burn-in sample of 100 observations in order to avoid initial value issues.

Table 9: Rejection probabilities of null hypotheses associated with the Mincer-Zarnowitz regression $r_t^2 = a + b(\hat{r}_{mt}^2 + \hat{\sigma}_{mt}^2) + u_{mt}$, using a nominal level of 10%, when $\mathbf{a}$ is equal to the benchmark values $(5^{-1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$

Model	T	Test 1	Test 2	Test 3	Test 4
		$H_0 : a = 0$	$H_0 : b = 0$	$H_0 : b = 1$	$H_0 : a = 0, b = 1$
		$H_1 : a \neq 0$	$H_1 : b \neq 0$	$H_1 : b \neq 1$	$H_1 : a \neq 0, b \neq 1$
1	25	0.22	0.44	0.33	0.36
	50	0.23	0.67	0.27	0.28
	100	0.17	0.91	0.19	0.21
	500	0.13	1.00	0.15	0.13
	1000	0.11	1.00	0.10	0.11
2	25	0.27	0.13	0.26	0.30
	50	0.36	0.19	0.23	0.52
	100	0.48	0.37	0.20	0.85
	500	0.77	0.97	0.14	1.00
	1000	0.88	1.00	0.11	1.00
3	25	0.32	0.15	0.42	0.46
	50	0.37	0.14	0.43	0.40
	100	0.38	0.17	0.40	0.33
	500	0.46	0.73	0.46	0.36
	1000	0.54	0.95	0.52	0.45
4	25	–	1.00	0.64	–
	50	–	1.00	0.71	–
	100	–	1.00	0.79	–
	500	–	1.00	0.99	–
	1000	–	1.00	1.00	–

Note: Simulations are in EViews 6 with 1000 replications, each with a prior burn-in sample of 100 observations in order to avoid initial value issues. White (1980) standard errors are used in all tests. The coefficient tests of $a = 0$ and $b = 0$ in tests 1 and 2 are two-sided, and the Wald coefficient restriction tests in tests 3 and 4 are the χ^2 versions.

Table 10: Rejection probabilities of the modified Diebold-Mariano (MDM), reality check (RC) and superior predictive ability (SPA) tests with a nominal level of 10% using MSE, MAE and residual kurtosis (Kurt) as loss functions when $\mathbf{a}$ is equal to the benchmark values $(5^{-1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$

T	m_4	MDM			T	Loss	RC	SPA
	vs.	MSE	MAE	Kurt				
25	m_1	0.38	0.61	0.00	25	MSE	0.25	0.23
	m_2	0.25	0.51	0.00		MAE	0.41	0.43
	m_3	0.24	0.53	0.00		Kurt	0.09	0.12
50	m_1	0.46	0.78	0.00	50	MSE	0.26	0.21
	m_2	0.30	0.72	0.00		MAE	0.47	0.50
	m_3	0.30	0.74	0.00		Kurt	0.08	0.08
100	m_1	0.62	0.95	0.00	100	MSE	0.33	0.22
	m_2	0.40	0.91	0.00		MAE	0.52	0.55
	m_3	0.37	0.92	0.00		Kurt	0.09	0.09
500	m_1	1.00	1.00	0.00	500	MSE	0.85	0.75
	m_2	0.91	1.00	0.00		MAE	0.90	0.91
	m_3	0.85	1.00	0.00		Kurt	0.31	0.31
1000	m_1	1.00	1.00	0.00	1000	MSE	0.98	0.96
	m_2	1.00	1.00	0.00		MAE	0.99	0.99
	m_3	0.98	1.00	0.00		Kurt	0.49	0.49

Note: Simulations are in R 2.6.1 and Ox 5/SPA 2.02 (see Hansen and Lunde 2007) with 1000 replications, each with a prior burn-in sample of 100 observations in order to avoid initial value issues. The MDM test uses a $t(1)$-distribution for the test-statistic, and in the RC and SPA simulations the nominal value is compared with the consistent p-value. All three tests are one-sided, and the number of bootstraps and the value of the dependence parameter in the RC and SPA tests are 1000 and 0.5, respectively.

Table 11: *Ex post* out-of-sample evaluation of 1-step weekly Norwegian exchange rate variability forecasts 26 January 2007 - 4 January 2008 (50 observations)

	MAE	MDM	a	b	R^2	AR_1	$\chi^2(2)$
ECON	0.59	0.67	-0.50	1.81	0.48	0.14	1.52
	[0.13]	[0.31]	[0.24]	[0.02]		[0.70]	[0.47]
RV	0.84	-1.30	-0.35	1.01	0.20	2.99	15.71
		[0.79]	[0.42]	[0.07]		[0.08]	[0.00]
EGARCH(1,1)	0.81	-1.57	0.89	-0.29	0.01	0.50	80.03
		[0.82]	[0.00]	[0.05]		[0.48]	[0.00]
Constant	0.67	–	–	1.30	0.00	0.73	–
				[0.00]		[0.39]	

Note: The first column contains the variability forecast MAE for each model, where the variability forecast error of model m at t is defined as $(r_t^2 - \hat{r}_{mt}^2 - \hat{\sigma}_{mt})$, and the consistent p-value of Hansen's (2005) SPA test in square parentheses. Column two contains MDM tests against the constant model as benchmark, columns three and four contain the OLS estimated parameter estimates of the Mincer-Zarnowitz regression $r_t^2 = a + b(\hat{r}_{mt}^2 + \hat{\sigma}_{mt}^2) + u_{mt}$, column five the associated R^2 of the regression, column six the Ljung and Box (1979) test-statistic (Q-stat.) for first order serial correlation in the residuals, and column seven contains the Wald test-statistic (χ^2 version) of a joint coefficient restriction test with $a = 0, b = 1$ as the null hypothesis using White (1980) estimates of the standard errors. Computations are in EViews 6, R 2.6.1 and Ox 5/SPA 2.02.

Table 12: *Ex ante* out-of-sample evaluation of 1-step weekly Norwegian exchange rate variability forecasts 26 January 2007 - 4 January 2008 (50 observations)

	MAE	MDM	a	b	R^2	AR_1	$\chi^2(2)$
ECON	0.74	-1.29	-0.37	0.36	0.01	0.10	46.96
	[0.65]	[0.79]	[0.41]	[0.55]		[0.80]	[0.00]
QV	0.87	-2.23	-0.25	0.22	0.01	0.48	42.36
		[0.87]	[0.33]	[0.56]		[0.49]	[0.00]
EGARCH(1,1)	0.81	-1.57	0.89	-0.29	0.01	0.50	80.03
		[0.82]	[0.00]	[0.05]		[0.48]	[0.00]
Constant	0.67	–	–	1.30	0.00	0.73	–
				[0.00]		[0.39]	

Note: The first column contains the variability forecast MAE for each model, where the variability forecast error of model m at t is defined as $(r_t^2 - \hat{r}_{mt}^2 - \hat{\sigma}_{mt})$, and the consistent p-value of Hansen's (2005) SPA test in square parentheses. Column two contains MDM tests against the constant model as benchmark, columns three and four contain the OLS estimated parameter estimates of the Mincer-Zarnowitz regression $r_t^2 = a + b(\hat{r}_{mt}^2 + \hat{\sigma}_{mt}^2) + u_{mt}$, column five the associated R^2 of the regression, column six the Ljung and Box (1979) test-statistic (Q-stat.) for first order serial correlation in the residuals, and column seven contains the Wald test-statistic (χ^2 version) of a joint coefficient restriction test with $a = 0, b = 1$ as the null hypothesis using White (1980) estimates of the standard errors. Computations are in EViews 6, R 2.6.1 and Ox 5/SPA 2.02.