Forecast Evaluation of Explanatory Models of Financial Variability[*]

Genaro Sucarrat[†]

4 March 2009

Abstract

A practice that has become widespread and widely endorsed is that of evaluating forecasts of financial variability obtained from discrete time models by comparing them with high-frequency *ex post* estimates (e.g. realised volatility) based on continuous time theory. In explanatory financial variability modelling this raises several methodological and practical issues, which suggests an alternative approach 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 variability are studied. Second, based on the simulation results a simple but general framework is proposed and illustrated. The illustration provides an example of where an explanatory model outperforms realised volatility *ex post.*

JEL Classification: C52, C53, F31, F37, F47
Keywords: Financial variability, financial volatility, forecasting, explanatory modelling, exchange rates

[*] I am indebted to various people for useful comments, suggestions and questions at different stages, including Álvaro Escribano, Katarina Juselius, Kristian Miltersen, Enrique Sentana, anonymous reviewers, seminar participants at the University of Oslo (September 2008), and participants at various conferences and workshops: Fibe 2009 (Bergen, January), Simposio de Análisis Económico 2008 (Zaragoza, December), Foro de Finanzas 2008 (Barcelona, November), The International Workshop on Recent Advances in Time Series Analysis (Protaras, June 2008), Inference and Tests in Econometrics—A Tribute to Russell Davidson (Marseille, April 2008), Forecasting Financial Markets 2008 (Aix-en-Provence, May) and The International Symposium of Forecasting 2007 (New York, June). 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.

[†] Department of Economics, Universidad Carlos III de Madrid (Spain). Email: gsucarra@eco.uc3m.es. Webpage: http://www.eco.uc3m.es/sucarrat/index.html

1 Introduction

The distinction between explanatory models on the one hand and non-explanatory models on the other is a matter of degree. An example of a model-class that is closer to the non-explanatory end of the continuum is "pure" time series models. That is, models that only contain an autoregressive moving average (ARMA) specification in the mean and/or an autoregressive conditional heteroscedasticity (ARCH) specification in the variance. Models in this class have proven to be of great value in financial *ex ante* forecasting, but their usefulness for conditional forecasting, impact analysis, counterfactual analysis, and scenario analysis more generally is severely limited in at least two ways. First, even though properties like ARMA and ARCH can be given economic interpretations, they remain silent about the specific economic phenomena that explains the property in question. For example, an ARCH model does not give an economic answer to the question *why* large (in absolute value) financial returns tend to cluster together. Indeed, for some policy-making purposes evidence of ARCH is useless without an adequate economic explanation of why. Second, the explanatory capacity of pure time series models is limited. For example, explanatory models can explain a substantial part of the remainder, that is, the error term of the pure time series model, if the right explanatory information is available.

Explanatory models of financial variability can thus provide essential insight beyond that of pure time series models in a wide range of situations.[1] 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 policy-making explanatory models can be used to inform policy decisions: They 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 evaluation plays an informative role in the assessment of explanatory models intended for any of these purposes.

A practice that has become widespread and widely endorsed is that of evaluating forecasts of financial variability obtained from discrete time models by comparing them with non-explanatory high-frequency *ex post* 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—or one of its related cousins either as comparison benchmark or as a measure of "true" inter-period volatility, where inter-period volatility is commonly defined

[1] Indeed, several commentators have even argued that some of the recent financial troubles could have been avoided had explanatory models been used to a greater extent instead of pure time series models.

as the conditional variance of financial inter-period return. The main motivation for the use of estimates made up of high-frequency intra-period data is that they are assumed to be more efficient estimates of volatility. In explanatory financial modelling, however, this approach raises several methodological and practical issues in addition to the limitations of non-explanatory models pointed out above. Below a non-exhaustive and incomplete list of five methodological and practical issues are given. For expository purposes the explanation of each issue is brief, but the first three of them are elaborated upon in greater detail in the appendix for the interested reader. The five issues are:

1. In empirical modelling both the mean residuals and the standardised residuals are derived in the sense that their properties depend on the functional form, and on the explanatory information in the mean and variance specifications. Since explanatory information typically is less available at high frequencies, explanatory low-frequency models can produce substantially better estimates of variability than high frequency models due to differing information sets. So evaluation procedures that treat neither type of models as more basic *a priori* are needed.

2. Since time is needed for an event to bring about another event, explanatory variables are likely to account for a decreasing portion of variability as the time increment goes to zero. Indeed, for philosophical reasons the portion will be equal to zero before the time increment reaches zero. A large part of modern finance theory is based on the idea that private information disseminates sequentially, and that it therefore aggregates temporally. In particular, it has been shown that order imbalance or order flow, a measure of temporally aggregated information, 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), Escribano and Pascual (2006), and Moberg (2008). However, as the time increment of a mathematical model goes to zero, the possibility of capturing such temporal aggregation effects vanishes. Consequently, no continuous time structure is capable of accounting for the whole range of possible temporal aggregation effects in an internally consistent manner. So even when explanatory information is available at high frequencies, explanatory modelling at lower frequencies can produce substantially different results compared with those implied by continuous time structures, because of temporal aggregation issues (see also Bertsimas et al. (2000) for a related argument).

3. 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.[2] In particular, continuous time structures implies

[2] This issue is analogous to the restrictions imposed by microfoundationalism: A macro model is always compatible with (in the sense that it can be derived from) more than one micro model.

excessive stability restrictions. For example, if a certain discretisation of a continuous time model is stable over time, then this does not imply that the underlying continuous time model is stable. Indeed, if *any* of the infinite discretisations is unstable, then—strictly speaking—the assumed underlying continuous time model is invalid. In practice one would not require that all discretisations of a continuous time model are stable.[3] Nevertheless, the excessive stability requirements only reiterates the need for procedures that enables us to evaluate discrete time models against estimates based on continuous time theory, without treating either as more fundamental.

4. It is well-known that measurement errors and a range of other market microstructure issues affect—possibly in substantial ways—the precision of high-frequency estimates, see amongst others Meddahi (2002), Barndorff-Nielsen and Shephard (2002) Aït-Sahalia and Mykland (2003), Andersen et al. (2005), Aït-Sahalia et al. (2005) and Aït-Sahalia (2007). Several adjustment procedures of the high-frequency estimates have been suggested in order to account for the presence of microstructure issues. However, it is not given *a priori* that the chosen error-adjusting method yield estimates that are better than those obtained from low-frequency models, possibly with explanatory information in the mean and variance specification (or both). So procedures that enable us to evaluate the estimates against each other without assuming *a priori* that any of the methods are more correct are needed.

5. The right combination of information and functional form in the conditional mean specification can result in homoscedastic (mean) errors. For example, an explicit aim of the General-to-Specific (GETS) methodology is to specify the conditional mean such that the mean errors become homoscedastic, since heteroscedasticity frequently is an indication of inadequate specification 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.

These methodological and practical issues suggest it is inappropriate to evaluate explanatory models' forecasts of variability by comparing them with high-frequency estimates of continuous time analogues. Instead, a natural and intuitively straightforward alternative that suggests itself is to compare variability forecasts in terms of predictions of squared financial return.[4] Squared return is an unsigned, direct and observable measure of the total relative gain or loss in the price of a financial

[3] In practice one would typically resort to the *ad hoc* assumption that the discretisation of interest is stable. Or, alternatively, in many cases a proportion equal to $1 - \alpha$, where α is the chosen nominal level of the stability test(s) in question under the null of stability, will do.

[4] I make no claim to originality in proposing squared returns as a measure of financial variability. Indeed, casual reading suggests it used to be one of the more common measures in academic finance and economics until the end of the 1990s.

asset, a magnitude which most economic and financial agents can relate to. By contrast, only agents that are (substantially) active in derivative markets are likely to have their profits and losses primarily dependent on volatility—a prediction of variability—rather than on variability itself. Another advantage of defining financial variability as squared return is that it then becomes an objectively given magnitude. In comparison, continuous time notions such as the integrated variance, which is an example of a continuous time analogue of volatility (realised volatility can be viewed as an estimate of integrated variance), is neither observable nor entirely objective, since its properties depend entirely on the assumptions of the assumed continuous time model. An equally straightforward, direct and objective measure of financial variability is absolute return. However, the advantage with squared return is that it more readily enables the joint modelling and analysis of level-effects (for example through error-correction terms), effects on return via the mean specification, and effects on volatility via the variance specification.

The main reason squared return fell out of favour during the 1990s is that it is considered an inefficient estimate of volatility, see for example Andersen and Bollerslev (1998). This naturally leads to the question: Is it feasible in practice to conceive of financial variability in terms of squared financial return? The purpose of this study is precisely to address this question by studying the finite sample properties of operational and practical evaluation procedures in a simulation study. Appropriate understanding in finite samples is crucial since explanatory data is typically available at lower frequencies only, say, daily, weekly, monthly, quarterly and yearly. Moreover, based on the simulation results I propose a general but simple framework that can be used to evaluate explanatory models, non-explanatory models, continuous time models and so on against each other without *a priori* assuming any of them as more basic.

The rest of the paper consists of three sections and one appendix. Section 2 contains the simulation study and the simple framework that is suggested with basis in the simulation results. Section 3 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 variability against high frequency estimates based on continuous time theory. Section 4 concludes and gives suggestions for further research. Finally, the appendix provides a more detailed characterisation of some of the methodological and practical issues that arise in the evaluation of explanatory models of financial variability.

2 Forecast evaluation of explanatory models of financial variability

Denote financial return in period t for r_t and consider the model

$$\begin{aligned} r_t &= \mu_t(b, x_t) + e_t & (1)\\ e_t &= \sigma_t z_t, \quad z_t \sim IID(0,1) & (2)\\ \sigma_t^2 &= h(\gamma, y_t). & (3) \end{aligned}$$

The first equation decomposes returns into a mean specification μ_t plus the mean error e_t, the second equation decomposes the mean error into the conditional standard deviation σ_t multiplied by an IID zero-mean and unit-variance process $\{z_t\}$ (that is, the standardised error or the "variance error"), and the third equation defines the conditional variance or volatility specification σ_t^2. The b and γ are parameter vectors, whereas x_t and y_t are vectors of conditioning variables that may include contemporaneous and lagged explanatory variables, and possibly lags of r_t and z_t. Financial variability is simply defined as r_t^2. Now, denote by $I_t = \{x_t, y_t\}$ the information at t, and denote by $\mathcal{I}_t = \{I_t, I_{t-1}, \ldots\}$ the information up to and including t. The conditional variability and the conditional variance or volatility are defined as $E(r_t^2|\mathcal{I}_t) = \mu_t^2 + \sigma_t^2$ and $Var(r_t|\mathcal{I}_t) = \sigma_t^2$, respectively. In other words, when the mean specification μ_t equals zero, then the conditional variability $E(r_t^2|\mathcal{I}_t)$ and the conditional variance or volatility $Var(r_t|\mathcal{I}_t)$ coincide.

Econometric models like (1)-(3) are simplified and partial representations of a highly complex and evolving social reality, and the probabilistic study of their relation belongs to econometric reduction theory, see amongst others Hendry and Richard (1982), Florens et al. (1990), Hendry (1995, chapter 9), Spanos (1999), Davidson (2000) and Sucarrat (2009). An important implication of econometric reduction theory is that the properties of the errors e_t and z_t are derived in the sense that they depend on the conditioning vectors x_t and y_t, and on how the conditioning information is used in the mean and variance specifications μ_t and σ_t^2 (see appendix for a more detailed discussion). As a consequence, evaluating discrete models by comparing their conditional forecasts $E(r_t^2|\mathcal{I}_t)$ of variability r_t^2 with high-frequency estimates of continuous time analogues of volatility σ_t^2 can be misleading, in particular when the explanatory information in the mean or variance specification (or both) has notable explanatory power. The purpose of this section is to study, by means of a simulation study, the properties of some procedures that enable us to evaluate discrete time and continuous time models against each other without treating either as more basic *a priori*. In particular, two questions are addressed: (1) What is the most appropriate loss function?, and (2) What is the most appropriate out-of-sample forecast test? A substantial number of studies have contributed to the understanding of these questions within the paradigm of volatility being *given* and independent of the modeller (as opposed to *determined* by the explanatory information included and the functional form chosen by the investigator), see amongst others Andersen and Bollerslev (1998), Meddahi (2002), Andersen et al. (2005), Hansen and Lunde (2005, 2006), and Patton (2007). So it is worth re-iterating that, 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, and as a result of how the information is used (functional form) by the investigator. Consequently, the magnitude to forecast is variability r_t^2, and volatility σ_t^2 can be considered an appropriate forecast of variability when the mean μ_t is equal to or approximately equal to zero.

Subsection 1 motivates and describes the simulation setup, subsection 2 motivates and describes the comparison models, subsection 3 sheds light on what the most appropriate loss function is, subsections 4 and 5 study the appropriateness of some common out-of-sample forecast tests, and subsection 6 outlines a general but simple framework for out-of-sample return variability comparison, which is to be illustrated in section 4 with a real data set.

2.1 Simulation setup

In order to understand the generality (or lack thereof) of the results of any simulation study, it is useful to distinguish between the actual DGP on the one hand and a simulation DGP on the other. The former is the actual process that generates the data, whereas the latter is at best a statistically valid representation of the actual DGP. In other words, the results of any simulation study applies, strictly speaking, only when the simulation DGP is a valid or approximately valid representation of the actual DGP.

The simulation DGP is given by

$$\begin{aligned} r_t &= bx_t + e_t, \quad e_t = \sigma_t z_t, \quad x_t \sim IIN(0,1), \quad z_t \sim IIN(0,1), \\ \sigma_t^2 &= \omega + \alpha e_{t-1}^2 + \beta\sigma_{t-1}^2 + cy_t, \quad y_t \sim IID, \quad y_t \in \{0,1\} \text{ with } P(1) = p, \end{aligned} \tag{4}$$

for $t = 1, \dots, T$, where x_t, z_t and y_t are mutually independent for all t. Although seemingly simple, the simulation DGP can actually be viewed as approximating a wide range of explanatory models of financial asset prices, and many models of the autoregressive conditional heteroscedasticity (ARCH) and stochastic volatility (SV) classes. The term bx_t is the explained portion of conditional first moment return variation and may be interpreted as approximating a structure that contains (say) contemporaneous and/or lagged money market variables ("interest rates"), stock market variables, order flow variables, news variables, and so on. For the purpose of a specific example, bx_t may be seen as approximating (say) $b_0 + b_1\Delta of_t + b_2\Delta ir_t + b_3\Delta p_t + b_4 ECM(s_{t-1}, of_{t-1}, ir_{t-1}, p_{t-1})$, where $r_t = \Delta s_t$, where of_t, ir_t and p_t denote cumulative order flow, the domestic inter-bank offer rate (a money market interest rate) and an external financial asset price, respectively, and where $ECM(\cdot)$ is an error-correction mechanism. A similar interpretation applies to the variance

Table 1: Descriptive statistics of interdaily (close, weekends excluded, $T = 594$) and weekly (close, Friday-to-Friday, $T = 118$) exchange rate returns in percent from 26 September 2005 to 4 January 2008

		USD/EUR	YEN/EUR	GBP/EUR	NOK/EUR
Daily:	$S.E.$	0.446	0.564	0.313	0.356
	$Kurtosis$	3.741	4.918	3.571	3.941
	JB	14.825	118.750	18.965	25.487
		[0.00]	[0.00]	[0.00]	[0.00]
	$AR(1)$	0.422	4.003	1.984	1.886
		[0.52]	[0.05]	[0.16]	[0.17]
	$ARCH(1)$	0.406	56.119	0.938	3.557
		[0.52]	[0.00]	[0.33]	[0.06]
Weekly:	$S.E.$	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]

$S.E.$ 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.

specification. 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 covariance non-stationarity). However, the GARCH(1,1) term can be viewed as approximating a structure that nests a wide range of explanatory models of volatility persistence, for example models that contain volume and/or other liquidity variables. The last term cy_t in the variance specification is a Bernoulli jump process, that 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. The term cy_t can therefore be viewed as approximating explanatory models of non-persistent volatility that contains, say, contemporaneous and/or lagged news and/or unexpected events, shocks, and so on.

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 (4) 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 accounts for an equal portion of total unconditional variability. In other words, in the benchmark simulation 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 benchmark 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 $b = 0$, $c = 0$ and $(\omega, \alpha, \beta) = (0.02, 0.8, 0.1)$, then 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 in the benchmark simulation DGP is set to 0.1, which means there is a jump once 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 (4) with $\mathbf{a} = (5^{-1/2}, 0.02, 0.1, 0.8, 2, 0.1)$.

Table 2 contains descriptive statistics of simulated returns for different values of $\mathbf{a}$, and compares with table 1 which contains descriptive statistics of the returns of four selected daily and weekly exchange rates from 30 September 2005 to 4 January 2008. The four exchange rates are USD/EUR, YEN/EUR, GBP/EUR and NOK/EUR. The first three are the most commonly traded currency pairs involving the euro, and compares with the NOK/EUR which will be used in the empirical illustration in the next section. The choice of NOK/EUR and sample period for the empirical illustration is partly due to the fact that microstructure issues are more likely to affect the NOK/EUR exchange rate since it is traded less, and partly because explanatory data with high explanatory power (order flow) is readily available for the NOK/EUR exchange rate. Nevertheless, it should be noted that the "stylised" properties of the NOK/EUR exchange rate returns reported in table 2 do not suggest that the NOK/EUR behaves fundamentally differently from the other more liquid currency pairs.

For the benchmark values $\mathbf{a}_1 = (5^{-1/2}, 0.02, 0.1, 0.8, 2, 0.1)$ the simulated standard error is 0.77. This is higher than the four daily standard errors and about the same as the lower of the two weekly standard errors. Removing the jump term, the

Table 2: Descriptive statistics of simulated returns for parameter values $\mathbf{a}_l$ ($l = 1, \ldots, 5$) when the simulation DGP is given by (4)

	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_4$	$\mathbf{a}_5$
$S.E.$	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

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 $\mathbf{a}_1 = (5^{-1/2}, 0.02, 0.1, 0.8, 0.2, 0.1)$, that is, the benchmark values, $\mathbf{a}_2 = (5^{-1/2}, 0.02, 0.1, 0.8, 0, 0)$, $\mathbf{a}_3 = (5^{-1/2}, 0.2, 0, 0, 0, 0)$, $\mathbf{a}_4 = (0, 0.02, 0.1, 0.8, 0, 0)$ 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 $r_{lt} = \hat{b}_0 + \hat{b}_1 x_t + \hat{e}_{lt}$, and R^2 *variability* is the average R^2 of the OLS regression $r_{lt}^2 = \hat{a}_0 + \hat{b}_1(\hat{r}_{lt}^2 + \hat{\sigma}_{lt}^2) + \hat{u}_{lt}$, where $(\hat{r}_{lt}^2 + \hat{\sigma}_{lt}^2)$ is conditional variability for $\mathbf{a}_l$.

persistence term and the mean 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 daily standard errors of GBP/EUR and NOK/EUR returns. 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 exhibit lower kurtosis among the eight empirical estimates. Four of the remaining six empirical kurtosis values 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. Graphic inspection suggests the high kurtosis of YEN/EUR returns is 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 2 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 has a large impact on the explanatory power of x_t, and that the persistence terms do not have notable impact on the explanatory power with respect to the benchmark simulation DGP. 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 2 contains the R^2 of the OLS estimated 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 under $\mathbf{a}_l$. It is commonly found that these so-called Mincer and Zarnowitz (1969) regressions of r_t^2 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.

2.2 Comparison models

Four models will be studied and compared given (4) as the DGP. At each simulation the "correct" parameter values for each specification will be used instead of estimated parameter estimates, because estimation at each simulation raises additional issues that would have to be addressed very carefully.[5] In other words, the simulation study may be seen as replicating situations where the estimation algorithm is reasonably successful in identifying the "correct" estimates of the specification in question.

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

$$\hat{r}_{1t} = bx_t, \qquad \hat{\sigma}_{1t}^2 = \omega + \alpha e_{1t-1}^2 + \beta\sigma_{t-1}^2 + cy_t \tag{5}$$

where $e_{1t} = r_t - bx_t$, and where 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

$$\hat{r}_{2t} = 0, \qquad \hat{\sigma}_{2t}^2 = [\omega + b^2(1 - \alpha - \beta)] + \alpha e_{2t-1}^2 + \beta\sigma_{t-1}^2 + cy_t \tag{6}$$

where $e_{2t} = r_t$, and where 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 correct unconditional variability $E(r_t^2)$ in the limit. The third model is intended to mimic the situation where one fits a model that only includes persistence terms. 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

$$\hat{r}_{3t} = 0, \qquad \hat{\sigma}_{3t}^2 = [\omega + b^2(1 - \alpha - \beta) + cp] + \alpha e_{3t-1}^2 + \beta\sigma_{t-1}^2 \tag{7}$$

where $e_{3t} = r_t$, and where the model's standardised residual $\hat{z}_{3t}$ is defined as $r_t/\hat{\sigma}_{3t}$. Here the augmented constant in the variance specification is specified as $[\omega + b^2(1 - \alpha - \beta) + cp]$ in order to adjust for the zero mean specification and the absence of the

[5] For a more detailed discussion, see my response to point 5 raised by anonymous referee number 2 in the "Specific remarks" part during the public review process of the discussion paper version of this article: `http://www.economics-ejournal.org/economics/discussionpapers/2008-18/`.

jump term. Again, the motivation is to ensure that the unconditional variability of model 3 is equal to the correct 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

$$\hat{r}_{4t} = 0, \qquad \hat{\sigma}_{4t}^2 = \text{simulated sample variance of } r_t, \tag{8}$$

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

2.3 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? 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 three 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

$$MSE_m = \frac{1}{T}\sum_{t=1}^{T}(r_t^2 - \hat{r}_{mt}^2 - \hat{\sigma}_{mt}^2)^2. \tag{9}$$

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

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

The lower the MAE_m, the greater proportion of variability is explained by model m. Finally, the third type of loss function that will be studied is the multiple correlation coefficient R^2 of regressions of the type

$$r_t^2 = a + b(\hat{r}_{mt}^2 + \hat{\sigma}_{mt}^2) + u_{mt}, \quad t = 1, \ldots, T, \tag{11}$$

Table 3: Probabilities of obtaining a correct ranking of models 1 to 4 using MSE, MAE 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	R^2
25	0.28	0.44	0.33
50	0.39	0.49	0.47
100	0.52	0.57	0.65
500	0.85	0.62	0.96
1000	0.96	0.61	0.99

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.

where a and b are parameters, and where u_t is the error term. 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 3 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. This is indeed the case for MSE and R^2, but not always the case for MAE. Comparing MSE and R^2, the latter is more likely (between 3 and 13 percentage points) to provide the correct ranking at all the studied sample sizes. It should be noted though 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 generally is preferable to R^2, since a biased model can produce high R^2 although far off. The MAE increases 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 4). The probabilities of correctly ranking the other models always increase with sample size when MAE is used, but not for model 4. Another result that is clear from the simulations, and which is of practical interest, is that in small samples, say, (approximately) when $T \leq 100$, then the MAE is considerably more likely to provide a correct ranking than MSE, and the magnitude is sufficiently high to be of practical use. Additional simulations (not reported) with different parameter values predictably suggest that the probabilities in table 3 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 characteristic 3 in the next paragraph).

Table 4 contains the simulated probabilities for each of the model's rank when the DGP is given by (4) and the benchmark values $\mathbf{a}_1$. There are at least three

Table 4: Ranking probabilities for models 1 and 2 using MSE, MAE 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	Model 1			Model 2			Model 3			Model 4		
		MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2
25	1st.	0.67	0.70	0.68	0.15	0.12	0.14	0.12	0.17	0.18	0.07	0.02	0.00
	2nd.	0.10	0.10	0.17	0.42	0.56	0.42	0.32	0.30	0.42	0.16	0.04	0.00
	3rd.	0.16	0.18	0.16	0.34	0.29	0.44	0.42	0.51	0.40	0.09	0.04	0.00
	4th.	0.08	0.03	0.00	0.09	0.04	0.00	0.15	0.02	0.00	0.69	0.91	1.00
50	1st.	0.78	0.80	0.86	0.12	0.10	0.08	0.07	0.10	0.06	0.03	0.01	0.00
	2nd.	0.07	0.08	0.07	0.50	0.57	0.52	0.32	0.34	0.42	0.12	0.02	0.00
	3rd.	0.12	0.12	0.08	0.32	0.30	0.40	0.48	0.55	0.52	0.08	0.04	0.00
	4th.	0.04	0.01	0.00	0.06	0.03	0.00	0.13	0.02	0.00	0.77	0.93	1.00
100	1st.	0.88	0.88	0.95	0.07	0.06	0.03	0.04	0.09	0.02	0.01	0.00	0.00
	2nd.	0.05	0.06	0.04	0.58	0.62	0.66	0.32	0.32	0.30	0.05	0.00	0.00
	3rd.	0.07	0.06	0.02	0.32	0.31	0.31	0.57	0.59	0.68	0.05	0.03	0.00
	4th.	0.01	0.00	0.00	0.03	0.01	0.00	0.07	0.00	0.00	0.89	0.97	1.00
500	1st.	1.00	1.00	1.00	0.00	0.10	0.00	0.00	0.04	0.00	0.00	0.00	0.00
	2nd.	0.00	0.00	0.00	0.85	0.66	0.96	0.14	0.34	0.04	0.00	0.00	0.00
	3rd.	0.00	0.00	0.00	0.14	0.24	0.04	0.85	0.63	0.96	0.00	0.10	0.00
	4th.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.90	1.00
1000	1st.	1.00	1.00	1.00	0.00	0.20	0.00	0.00	0.05	0.00	0.00	0.00	0.00
	2nd.	0.00	0.00	0.00	0.96	0.66	0.99	0.04	0.34	0.01	0.00	0.00	0.00
	3rd.	0.00	0.00	0.00	0.04	0.15	0.01	0.96	0.62	0.99	0.00	0.16	0.00
	4th.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.84	1.00

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.

characteristics of interest that emerge from the results:

1. 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 the 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 are 2% points more likely to incorrectly rank model 3 in front of model 2.

2. In samples smaller than (approximately) 100 observations, then the MAE is more likely than both MSE and R^2 to correctly rank each model regardless of the others' rank. In other words, the probability of correctly ranking a single model 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. 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%.

3. 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. The source of this anomaly is model 4 whose probability of being ranked correctly falls from 90% to 84% when T increases from 500 to 1000.

Additional simulations predictably suggest that the probabilities in table 4 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.

2.4 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 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) so-called "reality check" (RC) and Hansen's (2005) test for superior predictive ability (SPA) (cf. Hansen and Lunde 2005, and Bauwens and Sucarrat 2008). 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, $E[g(r_t, \hat{r}_t, \hat{\sigma}_t^2)] \leq E[g(r_t, \hat{r}_{mt}, \hat{\sigma}_{mt}^2)]$, even when the losses are possibly contemporaneously and/or serially correlated. Moreover, the simulations by Harvey et al. (1997) suggest the MDM test statistic behave quite well in samples of as small as eight observations. The RC and SPA tests can be viewed as variations of the MDM test, but differ in important ways. Instead of testing whether each of the comparison models is significantly better than the benchmark, they test whether the best model is significantly better or not, taking into account that the same data is re-used. That is, the RC and SPA tests control for the extent of "data mining". However, a disadvantage with RC and SPA is that they can be overly conservative (cf. Romano et al. 2008). 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 of this subsection is to assess the power of the MDM, RC and SPA tests under the alternative assumption that one or more models are better than the chosen benchmark. In the simulations we set model 4, the constant variability model, as the benchmark, which means that the three other models are better by construction. Two loss functions $g(\cdot)$ will be studied in the simulations, MSE and MAE. The loss differential d_t at t is defined as $g(r_t, \hat{r}_{4t}, \hat{\sigma}_{4t}^2) - g(r_t, \hat{r}_{mt}, \hat{\sigma}_{mt}^2)$ for $m = 1, 2, 3$. When MSE is the loss function then $d_t = (r_t^2 - \hat{r}_{4t}^2 - \hat{\sigma}_{4t}^2)^2 - (r_t^2 - \hat{r}_{mt}^2 - \hat{\sigma}_{mt}^2)^2$, where $r_t^2 - \hat{r}_{4t}^2 - \hat{\sigma}_{4t}^2$ is the variability forecast error of model 4 at t, and where $r_t^2 - \hat{r}_{mt}^2 - \hat{\sigma}_{mt}^2$ is the variability forecast error of model $m = 1, 2, 3$ at t. When MAE is the loss function then $d_t = |r_t^2 - \hat{r}_{4t}^2 - \hat{\sigma}_{4t}^2| - |r_t^2 - \hat{r}_{mt}^2 - \hat{\sigma}_{mt}^2|$.

Table 5 contains the simulated rejection probabilities of the null 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. 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 ideally 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 to or higher than the rejection probability of m_3 vs. m_4. Because multiple comparison tests are often used to choose among models, 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. 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

Table 5: 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 and MAE as loss functions when **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				
25	m_1	0.38	0.61	25	MSE	0.25	0.23
	m_2	0.25	0.51		MAE	0.41	0.43
	m_3	0.24	0.53				
				50	MSE	0.26	0.21
50	m_1	0.46	0.78		MAE	0.47	0.50
	m_2	0.30	0.72				
	m_3	0.30	0.74	100	MSE	0.33	0.22
					MAE	0.52	0.55
100	m_1	0.62	0.95				
	m_2	0.40	0.91	500	MSE	0.85	0.75
	m_3	0.37	0.92		MAE	0.90	0.91
500	m_1	1.00	1.00	1000	MSE	0.98	0.96
	m_2	0.91	1.00		MAE	0.99	0.99
	m_3	0.85	1.00				
1000	m_1	1.00	1.00				
	m_2	1.00	1.00				
	m_3	0.98	1.00				

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.

property. A third property that a multiple comparison test should ideally exhibit is of sufficiently high power to reject the null in small samples. The table suggests that this is indeed 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%, 26%, 30%, 53% and 73% for the five sample size ($T = 25, 50, 100, 500, 1000$). 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, $\mathbf{a} = (5^{-1/2}, 0.2, 0, 0, 0, 0)$, 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 5 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, (3) the power of the RC and SPA tests are relatively similar, since they differ a maximum of 10 percentage 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. But an even more serious shortcoming suggested by the additional simulations is that they do not provide clear guidance as to whether MSE or MAE is preferable, since their comparative power depends greatly on the parameter values of the DGP. Moreover, the additional simulations do not provide clear guidance as to whether RC or SPA is preferable nor under which circumstances. A likely reason for this is that both RC and SPA are asymptotic tests, whose properties are not necessarily approximated in (moderately) small samples. Presumably a more comprehensive and detailed simulation study could shed further light on these issues.

2.5 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

$$r_t^2 = a + b(\hat{r}_{mt}^2 + \hat{\sigma}_{mt}^2) + u_{mt}, \quad t = 1, \ldots, T \tag{12}$$

where $\hat{r}_{mt}^2 + \hat{\sigma}_{mt}^2$ is the variability forecast of model m. Ideally a and b should be equal or close 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 6 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 the constant variability forecast results in co-linearity between regressors). 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. Overall, the simulations suggest test 1 behaves as desired in large samples, but not in small samples. 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 is unlikely to be informative in practice in small samples.

Overall, 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 models 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 decrease 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 $b = 1$, and among the four tests this is the one that exhibits the most desirable properties. 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 characteristic that, for model 3, probabilities first fall until $T = 100$ and then increase. Overall, then, the results suggest that tests 1 and 4 can be useful in econometric practice, but the degree of usefulness depends

Table 6: 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	–

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.

on sample size. For small samples (T between 25 and 100) the tests may not be very informative, with test 1 being less informative than test 4. Finally, we can also conclude that Mincer-Zarnowitz regressions on constant variability predictions are not very useful, since tests 1 and 4 are not applicable to constant models of variability.

2.6 A simple framework

The simulations suggest the following simple but general three step framework for financial variability point forecast comparison:

1. 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. As a rule of thumb, MAE is more likely to provide the correct ranking in small samples of 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.

2. Compare 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. So although the RC and SPA tests provide additional information to the MDM, that additional information should be interpreted with great care.

3. Run 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.[6] The R^2 provides additional information on bias and how it can possibly be corrected. For example, if a model

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

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 by means of a linear transformation. Indeed, as we will see in the *ex post* comparison in the next section, an example of this is realised volatility.

3 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 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, the Norges Bank (The Central Bank of Norway) makes weekly order flow data of the Norwegian krone (NOK) freely available on their website, which means the explanatory power in terms of R^2 of exchange rate returns is likely to be relatively high.[7] Nevertheless, it should be noted that daily and weekly NOK/EUR returns do *not* behave in a fundamentally different way from the returns of more liquid currency pairs involving the euro, see table 1 in section 2.1.

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 using data from 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 expository 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

[7] The Norwegian order flow data are collected daily since 2 October 2005, but Norges Bank only makes weekly aggregates publicly available via their statistics webpages. Currently, the data can be downloaded via the url `http://www.norges-bank.no/templates/reportroot____60389.aspx` and are described in more detail in Meyer and Skjelvik (2006).

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.

3.1 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 (p-values in square brackets):

$$\begin{aligned}
100\Delta s_t &= \underset{[0.00]}{0.09}\Delta x_t - \underset{[0.02]}{1.48}(\Delta ir_t^{no} - \Delta ir_t^{emu}) - \underset{[0.00]}{17.11}ECM_{t-1} + \hat{e}_{1t}, \\
ECM_t &= s_t - 1.99 + 4.51 ir_t^{no} - 8.02 ir_t^{emu} \\
\hat{e}_{1t} &= \hat{\sigma}_{1t}\hat{z}_{1t}, \quad \hat{\sigma}_{1t} = 0.57, \quad \hat{z}_{1t} \sim IIN(0,1) \\
& R^2: \ 0.42 \quad AR_1: \ \underset{[0.31]}{1.02} \quad ARCH_1: \ \underset{[0.90]}{0.02} \quad JB: \ \underset{[0.99]}{0.03} \quad T = 68
\end{aligned}$$

The variable $100\Delta s_t$ 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 imply 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.[8] 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. 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

[8] The rawdata of s_t, ir_t^{no} and ir_t^{emu} are the daily series ew:nor19101, ew:14307 and ew:emu14813 from Reuters - EcoWin. The source of and further reading on the order flow data is contained in footnote 7.

associated with the tests. 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 $0.09\Delta x_t - 1.48(\Delta ir_t^{no} - \Delta ir_t^{emu}) - 17.11 ECM_{t-1}$, 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 forecasted. For simplicity, the *ex ante* forecast of the squared conditional mean forecast $\hat{r}_{1t}^2$ is specified as $(0.09)^2\overline{x} + (1.48)^2\overline{ir} + (17.11 ECM_{t-1})^2$, 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, including no measurement error and market microstructure noise, it can be shown that RV provides a consistent estimate of integrated variance (IV)—a continuous time analogue of discrete time volatility—when the time increment goes to zero (see appendix). 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 (2007) for an overview. For simplicity, however, since the forecast comparison is intended for illustration rather than as a comprehensive evaluation of state-of-the-art forecast models, only RV is included here. The weekly RV series is made up of 30-minute squared log-returns using end-of-interval mid-point quotes from Olsen Financial Technologies (OFT). The RV model is given by (p-values in square brackets):

$$\begin{aligned}
100\Delta s_t &= \hat{e}_{2t}, \quad \hat{e}_{2t} = \hat{\sigma}_{2t}\hat{z}_{2t}, \quad \hat{z}_{2t} \sim IIN(0,1) \\
\hat{\sigma}_{2t}^2 &= \sum_{n(t)=1}^{N(t)} (100\Delta s_{n(t)})^2 \\
& \quad R^2:\ 0.00 \quad AR_1:\ \underset{[0.80]}{0.06} \quad ARCH_1:\ \underset{[0.38]}{0.78} \quad JB:\ \underset{[0.68]}{0.77} \quad T = 68
\end{aligned}$$

The term $\hat{\sigma}_{2t}^2$ is RV at t and the diagnostic tests AR_1, $ARCH_1$ and JB are of the standardised residual $\hat{z}_{2t} = 100\Delta s_t / \hat{\sigma}_{2t}$. The *ex post* variability forecast is given by $\hat{\sigma}_{2t}^2$, that is, RV at t, whereas the *ex ante* forecast is given by the fitted values of an AR(1) model of RV.[9]

The third model is a plain exponential GARCH(1,1) model, that is, a plain EGARCH(1,1) model. The model is "plain" in the sense that the conditional mean

[9] 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.

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). Specifically the model is (p-values in square brackets):[10]

$$\begin{aligned} 100\Delta s_t &= \hat{e}_{3t}, \quad \hat{e}_{3t} = \hat{\sigma}_{3t}\hat{z}_{3t}, \quad \hat{z}_{3t} \sim IIN(0,1) \\ \log \hat{\sigma}^2_{3t} &= \underset{[0.00]}{-1.19} + \underset{[0.48]}{0.25}\left|\frac{\hat{e}_{3t-1}}{\hat{\sigma}_{3t-1}}\right| - \underset{[0.37]}{0.50}\log \hat{\sigma}^2_{3t-1} \\ & \qquad R^2:\ 0.00 \quad AR_1:\ \underset{[0.73]}{0.12} \quad ARCH_1:\ \underset{[0.80]}{0.06} \quad JB:\ \underset{[0.55]}{1.19} \quad T=68 \end{aligned}$$

The diagnostic tests are of the standardised residuals, and both the *ex post* and *ex ante* forecasts of variability are given by $\hat{\sigma}^2_{3t}$.[11]

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

$$\begin{aligned} \hat{\sigma}^2_4 &= \frac{1}{T}\sum_{t=1}^{T}(100\Delta s_t)^2 \\ & \quad R^2:\ 0.00 \quad AR_1:\ \underset{[0.99]}{0.00} \quad ARCH_1:\ \underset{[0.86]}{0.03} \quad JB:\ \underset{[0.49]}{1.44} \quad T=68 \end{aligned}$$

The diagnostic tests are of the standardised residual $\hat{z}_{4t} = r_t/\hat{\sigma}_4$, the square brackets contain the associated p-values, and both the *ex post* and *ex ante* forecasts are given by $\hat{\sigma}^2_4$.

3.2 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 the profit would 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 the purpose of conditional forecasting, counterfactual analysis and scenario analysis more generally. Table 7 contains the *ex post*

[10] It should be noted that we use a slightly reparametrised version of Nelson's (1991) model.

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

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, say, 1%, 5% and 10%. However, one should bear in mind that the 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. Also, the supporting simulations suggested the p-values of the SPA test should be handled with great care. 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 one should 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 relatively 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 straightforward manner.[12] The bias of the RV forecast nevertheless underlines that the precision of high-frequency estimates based on continuous time theory is questionable empirically. 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.[13] 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, counterfactual analysis and other situations where the assumption that the values of the conditioning variables are correct is appropriate. Explanatory models should ideally also be useful for forecasting when the values of the conditioning variables are uncertain. In such cases one cannot expect that explanatory models fare better than non-explanatory models. However, it is desirable that they fare at least *as well* as non-explanatory models. Table 7 contains the *ex ante* 1-step out-of-sample forecast evaluation results of the four models, and it should be noted that the respective *ex ante* forecasts of both the EGARCH(1,1) and constant models are equal to their respective *ex post* forecasts. 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 of the Mincer-Zarnowitz regressions are low and equal to 1% for ECON, RV and EGARCH(1,1), which is common in *ex ante* forecasting of vari-

[12] The RV also comes second after ECON according to MSE (not reported).

[13] 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.

Table 7: *Ex post* and *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	$\chi^2(2)$
Ex post:	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]	
Ex ante:	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]
	RV	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]	

The first column contains the variability forecast MAE for each model, where the variability forecast error of model m at t is defined as $(r_t^2 - \hat{r}_{mt}^2 - \hat{\sigma}_{mt}^2)$, 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.

ability. The joint test of $a = 0, b = 1$ in the Mincer-Zarnowitz regression, that is, the bias test, is rejected for all of the three models in which it can be undertaken. Overall, then, although the results do not point to a clear winner, the evidence is in favour of the constant variability model. Put differently, for the study of weekly NOK/EUR exchange rate variability 1-step ahead, the *ex post* results suggest ECON should be used in scenario analysis, say, conditional forecasting and counterfactual analysis, whereas the *ex ante* results suggest the constant model should be used in *ex ante* forecasting.

4 Conclusions

Evaluating explanatory models of financial inter-period return variability by comparing their forecasts with high-frequency intra-period estimates of continuous time analogues raises several methodological and practical issues. Together these methodological and practical issues suggest an alternative approach is needed, and this study has contributed in two ways. Firstly, the finite sample properties of operational and practical procedures have been studied, and the results suggest indeed that variability forecast comparison is feasible. Secondly, based on the simulation results, a simple but general framework has been proposed and illustrated.

The simple framework contains three steps. First, compute the MAE or MSE variability forecast errors where financial variability is defined as squared returns, and use the MAEs or MSEs to rank the models. Whether the MAE or MSE is more appropriate depends on sample size. As a rule of thumb, 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 forecast precision using the MDM and SPA/RC tests. The tests exhibit relatively high power when the benchmark values are used by the simulation DGP—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 variability, then the power can be very low—in particular in small samples. The third and final step consists of testing for forecast bias by means of a Mincer-Zarnowits regression of the actual value of variability (squared return) on a constant and the variability 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 to reflect the empirical properties financial returns actually exhibit, further study is needed. In particular, further investigation is needed in order to understand better how the loss functions and statistical tests behave when the standardised error is (much) 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.

References

Aït-Sahalia, Y. (2007). Estimating Continuous-Time Models with Discretely Sampled Data. In R. Blundell, T. Persson, and W. K. Newey (Eds.), *Advances in Economics and Econometrics, Theory and Applications, 9th. World Congress.* Cambridge: Cambridge University Press.

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 (2002). Estimating Quadratic Variation Using Realized Variance. *Journal of Applied Econometrics 17*, 457–477.

Bauwens, L. and G. Sucarrat (2008). General to Specific Modelling of Exchange Rate Volatility: A Forecast Evaluation. Forthcoming in the *International Journal of Forecasting*. UC3M Working Paper version: WP 08-18 in the Economic Series (available via `http://hdl.handle.net/10016/2591`).

Bertsimas, D., L. Kogan, and A. Lo (2000). When is time continuous? *Journal of Financial Economics 55*, 173–204.

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.

Davidson, J. (2000). *Econometric Theory*. Oxford: Blackwell Publishers Ltd.

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

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.-M. (2008). *Essays on Empirical Market Microstructure*. Ph. D. thesis, Department of Finance and Management Science, Norwegian School of Economics and Business Administration, Bergen, Norway.

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`.

Romano, J. P., A. Shaikh, and M. Wolf (2008). Formalized data snooping based on generalized error rates. *Econometric Theory 24*, 404–447.

Spanos, A. (1999). *Probability Theory and Statistical Inference.* Cambridge: Cambridge University Press.

Sucarrat, G. (2009). Econometric Reduction Theory and Philosophy. UC3M Working Paper 09-10 in the Economic Series. Available via `http://e-archivo.uc3m.es/`.

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

Tiles, M. (1989). *The Philosophy of Set Theory.* New York: Dover Publications, Inc.

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.

Appendix: Explanatory modelling of financial return variability

There is a curious twist to a current development in financial econometrics. On the one hand, continuous time methods are (uncritically) becoming more and more common and important. On the other hand, the widespread view among philosophers of time and among philosophers of mathematics is that mathematics in general and real numbers in particular are not capable of depicting continuous time in an entirely accurate manner. For many purposes the representation error incurred by mathematics is unimportant. However, when temporal aggregation issues are involved, as is often the case in financial econometrics, then it can lead to seriously erroneous conclusions. One of the problems stems from the so-called principle of extensionality, that is, the axiom that two sets (or elements of a sets) are equal if and only if they are the same. This axiom is needed by mathematics in order to avoid contradictions caused by some self-referential paradoxes, and the consequence of the axiom is essentially that mathematics is "discrete" and that the notion of continuity has to be approximated. Typically the axiom of infinity plays an important role in such approximations, and an example of a mathematical structure that is commonly used in order to approximate the idea of continuity is the set of real numbers.[14]

The focus of this appendix is on economic and econometric issues rather than on philosophy, although they of course are related. The purpose of this appendix is to provide a more detailed characterisation of the first three methodological and practical issues listed in the introduction. This characterisation is contained in the second and last part of this appendix. Before that, in the first part, a discussion of explanatory discrete time models as derived entities is needed. This discussion is useful in order to understand why volatility in explanatory models is not only a latent and unobservable variable to be estimated, but also an entity whose properties depend on functional form and on the explanatory power of the information in the conditional mean and variance specifications.

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), Spanos (1999), Davidson (2000, subsection 4.1) and Sucarrat (2009).[15] 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 $r_t = b_0 + b_1 x_t + e_t$. A shortcoming with Hendry's theory is that it cannot provide reduction analysis on the relation between continuous and discrete time models, since his theory is entirely couched in terms of discrete time variables. However, the non-restrictive

[14] See Tiles (1989) for an introduction to the philosophy of mathematics that takes continuity issues as its organising theme. Further reading on the philosophy of time and on the philosophy of mathematics can be found in most dictionaries and handbooks of philosophy (even Wikipeadia).

[15] 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.

modifications to the initial probability space in Hendry's theory proposed in Sucarrat (2009) enables reduction analysis on the relation between continuous and discrete time models.

A key implication of reduction theory is that the properties of an 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, I assume no measurement error in any of the variables of the DGP, so that the variables of the DGP correspond to that of the theory mechanism (cf. discussion in Sucarrat 2009). In this regard it should also be noted that volatility is *not* a variable in the DGP (nor in the theory mechanism), since the theory mechanism and the DGP are entities whose properties are independent of how we represent them by means of models.[16] Let the density $f(r_t, x_t, y_t)$ denote the DGP of r_t, x_t and y_t, where r_t is financial return, and where x_t and y_t are vectors of conditioning variables. Specifically, in addition to other contemporaneous and/or lagged explanatory variables the vectors x_t and y_t may also contain lags of r_t and/or transformations of lags of r_t. Suppose the discrete time representation

$$r_t = g(x_t, b) + e_t,$$

is a model of the conditional DGP given by $f(r_t|x_t)$ such that $g(x_t, b)$ is equal to the conditional mean $E[r_t|g(x_t, b)]$, where b is a parameter vector and e_t is the error term. The error term e_t is then defined as $r_t - g(x_t, b)$, and the properties of e_t are therefore derived or "designed" in the sense that they are a result of how $g(x_t, b)$ is specified subject to the conditional DGP given by $f(r_t|x_t)$.[17] In particular, the

[16] From a reduction theory point of view there is only one DGP, *the* DGP, and the DGP is the most accurate and complete probabilistic representation of reality. In other words, the DGP serves as a "probabilistic ontology" (a probabilistic representation of reality as it objectively is). The volatility models or continuous time structures that have been put forward in the literature can therefore not be a DGP in the reduction theoretical sense, since they are not accurate and/or complete enough to constitute a probabilistic representation of an ontology. However, they can constitute what I elsewhere call "estimation and inference" models (Sucarrat 2009, introduction and figure 1 in particular, but see also section 4). Moreover, since the DGP is intended to be something objective and independent of ourselves and our representations of it, unless you believe in a rather unusual version of Platonism, and Platonism itself is a questionable philosophical thesis, then volatility as such does not exist objectively. In empirical discrete time modelling this is straightforward: Volatility is simply a model of the unexplained portion of the mean, that is, a model of the error term, and volatility is therefore entirely determined by the modeller through the choice of specification, conditioning information, assumptions on the standardised residuals and so on. In (objective) continuous time, the thesis that continuous time (instantaneous) volatility exists independently and objectively would therefore have to resort to rather strong philosophical assumptions.

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

better $g(x_t, b)$ is specified and the more explanatory power carried by x_t, the smaller e_t is likely to be in absolute value.

Consider now the discrete time model

$$r_t = g(x_t, b) + e_t, \quad e_t = \sigma_t z_t, \tag{13}$$

$$\sigma_t^2 = h(y_t, c), \tag{14}$$

where c is a vector of parameters, and where σ_t^2 is discrete time volatility and equal to the conditional variance $Var[r_t|g(x_t, b), h(y_t, c)]$. The term $g(x_t, b)$ is now equal to the conditional mean $E[r_t|g(x_t, y_t)]$, and the standardised residual z_t is defined as $[r_t - g(x_t, b)]/\sigma_t$. The properties of z_t are therefore determined by the specification of $g(x_t, b)$ and $h(y_t, c)$ subject to the conditional DGP given by $f(r_t|x_t, y_t)$. In particular, the better $g(x_t, b)$ and $h(y_t, c)$ explain the variation in r_t and e_t^2, respectively, the smaller z_t is likely to be in absolute value.[18]

Continuous vs. discrete models

If models are entities that depend on the specification of the conditional mean and variance subject to the DGP, then discrete time volatility is not a *given* magnitude independent of the researcher as suggested by some scholars. On the contrary, the value and characteristics depend on the conditional mean and variance specifications, and the more so the better the explanatory variables explain the variation in return and in the squared error. Accordingly, evaluating volatility estimates from an explanatory discrete time model by comparing them with high-frequency estimates of continuous time analogues 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

$$r(t) = A(t) + M(t), \quad t \in [0, T], \tag{15}$$

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

$$r(t) = \int_{t-1}^{t} \mu(s)ds + \int_{t-1}^{t} \sigma(s)W(s)ds. \tag{16}$$

[18] Of course, as pointed out by the reviewers, the converse is not necessarily true: The smaller z_t is in absolute value, the better $g(x_t, b)$ and $h(y_t, c)$ explain the variation in r_t and e_t^2, respectively. The reason is that the $\{z_t\}$ can be made small in absolute value by setting $\{\sigma_t\}$ arbitrarily large.

In this particular case the integrated variance $\int_{t-1}^{t} \sigma(s)^2 ds$ serves as the counterpart of discrete time volatility σ_t^2 as defined in the discrete time model (13)-(14) above, and a common estimator of integrated variance is realised volatility, that is, the sum of equidistant intra-period squared returns.

Evaluating volatility estimates from the discrete time model (13)-(14) against estimates obtained based on, say, (16) raises several methodological and practical issues which were listed in the introduction. Here the first three of these issues are discussed in more detail:

1. Although financial price 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 (if any) fraction of the total variation in $r(t)$ when only high-frequency data are used for estimation. By contrast, when lower frequency data are used then $\int_{t-1}^{t} \mu(s)ds$ may account for a substantial fraction of the total return variation. A similar argument applies of course to values of $\int_{t-1}^{t} \sigma(s)ds$. If $\int_{t-1}^{t} \mu(s)ds$ is equal to or approximately equal to zero, then the value of $\int_{t-1}^{t} \sigma(s)ds$ is effectively determined by the assumptions regarding the process $\{W(t)\}$. By contrast, estimation of σ_t—the discrete time counterpart of $\int_{t-1}^{t} \sigma(s)ds$—can lead to substantially different values owing to explanatory information in either or both $g(x_t, b)$ and $h(y_t, c)$. As a consequence, the properties of the standardised residual z_t can be substantially different from the properties of $\int_{t-1}^{t} W(s)ds$. All this is not surprising since the two approaches use information sets that differ. But it nevertheless underlines the need for methods that enable us to evaluate discrete and continuous time models against each other without treating either as more basic.

2. Due to economic reasons both $A(t)$ and the "explanatory" component of $M(t)$—for example $\int_{t-1}^{t} \sigma(s)ds$ in (16)—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. Indeed, for philosophical reasons $A(t)$ will reach zero before the time increment reaches zero. 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.

3. Whenever it is assumed that a discrete time model can be derived from the continuous time model in question—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 (in a probabilistic sense) 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 with (and can thus be derived from) more than one continuous time structure. In terms of the concepts and terminology in Sucarrat (2009, see section 4 in particular), if $\mathcal{A} = \{A_1, A_2, \ldots\}$ are the sets of possible worlds in which the discrete time model (13)-(14) is "true", and if $\mathcal{B} = \{B_1, B_2, \ldots\}$ are the sets of possible worlds in which the continuous time model (15) is true, then the probabilities of (13)-(14) and (15), 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 (13)-(14) and (15) being true jointly, which is effectively the assumption upon which evaluation of discrete time estimates against continuous time estimates is based, is $P[(\bigcup_{i=1}^{\infty} A_i) \cap (\bigcup_{j=1}^{\infty} B_j)]$. Now, by the nature of probability it is always the case that $P[(\bigcup_{i=1}^{\infty} A_i) \cap (\bigcup_{j=1}^{\infty} B_j)] \leq P(\bigcup_{i=1}^{\infty} A_i)$. In words, the probability that both the discrete time model (13)-(14) and the continuous time model (15) are true is always equal to or smaller than the probability that only the discrete time model (13)-(14) is true. Another way to put this is that, in a probabilistic sense, it is not generally the case that continuous time models nest the discrete time models that can be derived from the former.