The Rimini Centre for Economic Analysis

WP 19-09

John M. Maheu
University of Toronto, Canada
and
The Rimini Centre of Economic Analysis

Thomas H. McCurdy
University of Toronto, Canada
and
CIRANO

DO HIGH-FREQUENCY MEASURES OF VOLATILITY IMPROVE FORECASTS OF RETURN DISTRIBUTIONS?

The Rimini Centre for Economic Analysis
Legal address: Via Angherà, 22 – Head office: Via Patara, 3 - 47900 Rimini (RN) – Italy
www.rcfea.org - secretary@rcfea.org

Do high-frequency measures of volatility improve forecasts of return distributions?

JOHN M. MAHEU and THOMAS H. MCCURDY*

First Draft: June 2006, This Draft: May 2009

Abstract

Many finance questions require the predictive distribution of returns. We propose a bivariate model of returns and realized volatility (RV), and explore which features of that time-series model contribute to superior density forecasts over horizons of 1 to 60 days out of sample. This term structure of density forecasts is used to investigate the importance of: the intraday information embodied in the daily RV estimates; the functional form for $\log(RV)$ dynamics; the timing of information availability; and the assumed distributions of both return and $\log(RV)$ innovations. We find that a joint model of returns and volatility that features two components for $\log(RV)$ provides a good fit to S&P 500 and IBM data, and is a significant improvement over an EGARCH model estimated from daily returns.

Keywords:

Realized Volatility, multiperiod out-of-sample prediction, term structure of density forecasts, Stochastic Volatility

*Maheu (jmaheu@chass.utoronto.ca), Department of Economics, University of Toronto and RCEA; McCurdy (tmccurdy@rotman.utoronto.ca), Rotman School of Management, University of Toronto, and CIRANO. We thank the editors and two anonymous referees, as well as Zhongfang He, Lars Stentoft, participants of the April 2006 CIREQ conference on Realized Volatility, the McGill 2008 Risk Management Conference, the University of Waterloo 2009 Econometrics and Risk Management Conference, and seminar participants at the Federal Reserve Bank of Atlanta for many helpful comments. Guangyu Fu and Xiaofei Zhao provided excellent research assistance. We are also grateful to the SSHRC for financial support.

1 Introduction

Many finance questions require a full characterization of the distribution of returns. Examples include option pricing which uses the forecast density of the underlying spot asset, or Value-at-Risk which focuses on a quantile of the forecasted distribution. Once we move away from the simplifying assumptions of Normally-distributed returns or quadratic utility, portfolio choice also requires a full specification of the return distribution.

The purpose of this paper is to study the accuracy of forecasts of return densities produced by alternative models. Specifically, we focus on the value that high frequency measures of volatility provide in characterizing the forecast density of returns. We propose new bivariate models of returns and realized volatility and explore which features of those time-series models contribute to superior density forecasts over multiperiod horizons out of sample.

Andersen and Bollerslev (1998), Andersen, Bollerslev, Diebold, and Labys (2001), Andersen, Bollerslev, Diebold, and Ebens (2001), Barndorff-Nielsen and Shephard (2002), and Meddahi (2002), among others,[1] have established the theoretical and empirical properties of the estimation of quadratic variation for a broad class of stochastic processes in finance. Although theoretical advances continue to be important, part of the research in this new field has focused on the time-series properties and forecast improvements that realized volatility provides. Examples include Andersen, Bollerslev, Diebold, and Labys (2003), Andersen, Bollerslev, and Diebold (2007), Andersen, Bollerslev, and Meddahi (2004), Ghysels and Sinko (2006), Ghysels, Santa-Clara, and Valkanov (2006), Koopman, Jungbacker, and Hol (2005), Maheu and McCurdy (2002, 2007), Martens, van Dijk, and de Pooter (2003), and Taylor and Xu (1997).

Few papers have studied the benefits of incorporating RV into the return distribution. Andersen, Bollerslev, Diebold, and Labys (2003), and Giot and Laurent (2004) consider the value of RV for forecasting and for Value-at-Risk. These approaches decouple the return and volatility dynamics and assume that RV is a sufficient statistic for the conditional variance of returns. Ghysels, Santa-Clara, and Valkanov (2005) find that high frequency measures of volatility identify a risk-return tradeoff at lower frequencies. Their filtering approach to volatility measurement

[1] Recent reviews include Andersen, Bollerslev, and Diebold (2009), Barndorff-Nielsen and Shephard (2007).

does not provide a law of motion for volatility and therefore multiperiod forecasts cannot be computed in that setting.

RV is an *ex post* measure of volatility and in general may not be equivalent to the conditional variance of returns. We propose bivariate models based on two alternative ways in which RV is linked to the conditional variance of returns. Since our system provides a law of motion for both return and RV at the daily frequency, multiperiod forecasts of returns and RV or the density of returns are available. The dynamics of the conditional distribution of RV will have a critical impact on the quality of the return density forecasts.

Our benchmark model is an EGARCH model of returns. This model is univariate in the sense that it is driven by one stochastic process which directs the innovations to daily returns. It does not allow higher-order moments of returns to be directed by a second stochastic process. Nor does it utilize any intraday information.

Two types of functional forms for the bivariate models of returns and RV are proposed. The first model uses a heterogeneous autoregressive (HAR) specification (Corsi (2009), Andersen, Bollerslev, and Diebold (2007)) of $\log(RV)$. A second model allows different components of $\log(RV)$ to have different decay rates (Maheu and McCurdy (2007)).

We also consider two ways to link RV to the variance of returns. First, we impose the cross-equation restriction that the conditional variance of daily returns is equal to the conditional expectation of daily RV. Second, motivated by Bollerslev, Kretschmer, Pigorsch, and Tauchen (2009) who model returns, bipower variation and realized jumps in a multivariate setting,[2] we also investigate a specification of our bivariate component model for which the variance of returns is assumed to be synonymous with RV. We label this case 'observable stochastic volatility' and explore whether this assumption improves the term structure of density forecasts. We also compare specifications with non-Normal versus Normal innovations for both returns and $\log(RV)$.

As in our benchmark EGARCH model, all of our bivariate models allow for so-called leverage or asymmetric effects of past negative versus positive return innovations. Our bivariate models allow for mean reversion in RV. This allows us to evaluate variance targeting for these

[2]For definition and development of bipower variation and realized jumps see, for example, Barndorff-Nielsen and Shephard (2004).

specifications.

Our main method of model comparison uses the predictive likelihood of returns. This is the forecast density of a model evaluated at the realized return; it provides a measure of the likelihood of the data being consistent with the model. Intuitively, better forecasting models will have higher predictive likelihood values. Therefore our focus is on the relative accuracy of the models in forecasting the return density out of sample. The forecast density of the models is not available in closed form; however, we discuss accurate simulation methods that can be used to evaluate the forecast density and the predictive likelihood.

An important feature of our approach is that we can directly compare traditional volatility specifications, such as EGARCH, with our bivariate models of return and RV since we focus on a common criteria – forecast densities of returns. We generate a predictive likelihood for each out-of-sample data point and for each forecast horizon. For each forecast horizon, we can compute the average predictive likelihood where the average is computed over the fixed number of out-of-sample data points. A term structure of these average predictive likelihoods allows us to investigate the relative contributions of RV over short to long forecast horizons.

Our empirical applications to S&P 500 (Spyder) and IBM returns reveal the importance of intraday return information, the timing of information availability, and non-Normal innovations to both returns and $\log(RV)$. The main features of our results are as follows. Bivariate models that use high frequency intraday data provide a significant improvement in density forecasts relative to an EGARCH model estimated from daily data. Two-component specifications for $\log(RV)$ provide similar or better performance than HAR alternatives; both dominate the less flexible single-component version. A bivariate model of returns with Normal innovations and observable stochastic volatility directed by a 2-component, exponentially decaying function of $\log(RV)$ provides good density forecasts over a range of out-of-sample horizons for both data series. We find that adding a mixture of Normals or GARCH effects to the innovations of the $\log(RV)$ part of this specification is not statistically important for our sample of S&P 500 returns, while the addition of the mixture of Normals provides a significant improvement for IBM.

This paper is organized as follows. The next section introduces the data used to construct daily returns and daily RV. It also discusses the measurement of volatility, the adjustments to realized volatility to remove the effects of market microstructure, and a benchmark model which is based on daily return data. Our bivariate models of returns and RV, based on high-frequency intraday data, are introduced in Section 3. The calculation of density forecasts and the predictive likelihood are discussed in Section 4; results are presented in Section 5. Section 6 concludes.

2 Data and Realized Volatility Estimation

We investigate a broadly diversified equity index (the S&P 500) and an individual stock (IBM). For the former we use the Standard & Poor's Depository Receipt (Spyder) which is a tradable security that represents ownership in the S&P 500 Index. Since this asset is actively traded, it avoids the stale price effect associated with using the S&P 500 index at high frequencies. Transaction price data associated with both the Spyder and IBM are obtained from the New York Stock Exchange's Trade and Quotes (TAQ) database.

Our data samples cover the period January 2, 1996 to August 29, 2007 for the Spyder and January 4, 1993 to August 29, 2007 for IBM. The shorter sample for the Spyder data was chosen based on volume of trading, for example there were many 5-minute periods with no transactions during the first years after the Spyder started trading in 1993, and a structural break in the Spyder $\log(RV)$ data in the mid 1990s (Liu and Maheu (2008)). The average number of transactions per day for the 1996-2007 sample of Spyder data was $32,971$ but the volume of trades has increased substantially over the sample – especially from 2005 forward. In contrast, the average number of transactions per day for IBM shares has been more stable over our 1993-2007 sample, averaging $6,011$ transactions per day with a substantial increase from late 2006.

After removing errors from the transaction data,[3] a 5-minute grid[4] from 9:30 to 16:00 EST was constructed by finding the closest transaction price before or equal to each grid-point time. From this grid, 5-minute continuously compounded (log) returns were constructed. These returns

[3]Data were collected with a TAQ correction indicator of 0 (regular trade) and when possible a 1 (trade later corrected). We also excluded any transaction with a sale condition of Z, which is a transaction reported on the tape out of time sequence, and with intervening trades between the trade time and the reported time on the tape. We also checked any price change that was larger than 3% and removed obvious errors.

[4]Volatility signature plots using grids ranging from 1 minute to 195 minutes are available on request.

were scaled by 100 and denoted as $r_{t,i}, i = 1, ..., I$, where I is the number of intraday returns in day t. For our 5-minute grid, normally $I = 78$ although the market closed early on a few days. This procedure generated 228,394 5-minute returns corresponding to 2936 trading days for the S&P 500; and 286,988 5-minute returns corresponding to 3693 trading days for IBM.

The increment of quadratic variation is a natural measure of *ex post* variance over a time interval. A popular estimator of it is realized variance or realized volatility (RV) computed as the sum of squared returns over this time interval. The asymptotic distribution of RV has been studied by Barndorff-Nielsen and Shephard (2002) who provide conditions under which RV is an unbiased estimate.

Given the intraday returns, $r_{t,i}, i = 1, ..., I$, an unadjusted daily RV estimator is

$$RV_{t,u} = \sum_{i=1}^{I} r_{t,i}^2. \tag{2.1}$$

However, in the presence of market-microstructure dynamics, RV can be a biased and inconsistent estimator for quadratic variation (Bandi and Russell (2008) and Zhang, Mykland, and Aït-Sahalia (2005)). Therefore, we consider several adjustments to our estimates and gauge their statistical performance in our model comparisons.[5]

Hansen and Lunde (2006) suggest the use of Bartlett weights to rule out negative values for RV. Following this approach, a corrected RV estimator is

$$RV_{t,ACq} = \omega_0 \hat{\gamma}_0 + 2 \sum_{j=1}^{q} \omega_j \hat{\gamma}_j, \quad \hat{\gamma}_j = \sum_{i=1}^{I-j} r_{t,i} r_{t,i+j}, \tag{2.2}$$

in which the weights follow a Bartlett scheme $\omega_j = 1 - \frac{j}{q+1}$, $j = 0, 1, ..., q$. We consider $q = 1, 2, 3$. Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) discuss the asymptotic properties of statistics of this type.

In order to match the volatility measures, daily returns, r_t, are computed as the logarithmic difference of the closing price and the opening price. These returns are scaled by 100. Table 1

[5] For alternative approaches to dealing with market microstructure dynamics see Aït-Sahalia, Mykland, and Zhang (2005), Bandi and Russell (2006), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008), Oomen (2005), Zhang (2006) and Zhou (1996).

displays summary statistics for daily returns and daily RV estimates computed from the 5-minute grid. If we take the sample variance of daily returns as a benchmark estimate of volatility in which no market microstructure effects are present, and compare this to the sample mean of RV, we see a clear bias for unadjusted RV. With respect to removing bias, it appears that a Bartlett adjustment with $q = 3$ is necessary for the S&P 500 (Spyder) data, whereas an adjustment with $q = 1$ is adequate for the IBM data. This conclusion is supported by autocorrelation analyses of the 5-minute returns data, as revealed by the autocorrelation functions with associated confidence bounds in Figure 1 for the S&P 500 and IBM respectively. For the remainder of our paper, unless otherwise stated, we use $RV_t \equiv RV_{t,ACq}$, with $q = 3$ for the S&P 500 and $q = 1$ for the IBM data.

One way to ascertain whether or not high-frequency (intraperiod) information contributes to improved forecasts of return distributions, is to compare density forecasts from our bivariate specifications of returns and $\log(RV)$ with those from a benchmark EGARCH specification:

$$r_t = \mu + \epsilon_t, \quad \epsilon_t = \sigma_t u_t \quad u_t \sim NID(0,1), \tag{2.3}$$

$$\log(\sigma_t^2) = \omega + \beta \log(\sigma_{t-1}^2) + \gamma u_{t-1} + \alpha |u_{t-1}|. \tag{2.4}$$

3 Joint Return-RV Models

As discussed in the Introduction, an integrated model of returns and realized volatility is needed to deal with common questions in finance which require a forecast density of returns for multiple horizons. In this section, we introduce two alternative joint specifications of daily returns and realized volatility. These bivariate models are distinguished by alternative assumptions about RV dynamics. We also consider versions of these bivariate models with non-Normal return and $\log(RV)$ innovations, as well as a version with an alternative assumption concerning available information about RV. In each case, cross-equation restrictions link the variance of returns and our realized volatility specification.

Corollary 1 of Andersen, Bollerslev, Diebold, and Labys (2003) shows that, under empirically realistic conditions, the conditional expectation of quadratic variation (QV_t) is equal to the conditional variance of returns, that is, $E_{t-1}(QV_t) = \text{Var}_{t-1}(r_t) \equiv \sigma_t^2$. If RV is an unbiased estimator of quadratic variation,[6] it follows that the conditional variance of returns can be linked to RV as $\sigma_t^2 = E_{t-1}(RV_t)$ where the information set is defined as $\Phi_{t-1} \equiv \{r_{t-1}, RV_{t-1}, r_{t-2}, RV_{t-2}, ..., r_1, RV_1\}$. Assuming that RV has a log-Normal distribution, that restriction takes the form

$$\sigma_t^2 = E_{t-1}(RV_t) = \exp\left(E_{t-1}\log(RV_t) + \frac{1}{2}\text{Var}_{t-1}(\log(RV_t))\right). \tag{3.1}$$

3.1 HAR-RV Specifications

We begin with a bivariate specification for daily returns and RV in which conditional returns are driven by Normal innovations and the dynamics of $\log(RV_t)$ are captured by Heterogeneous AutoRegressive (HAR) functions of lagged $\log(RV_t)$. Corsi (2009) and Andersen, Bollerslev, and Diebold (2007) use HAR functions in order to parsimoniously capture long-memory dependence. Motivated by that work, we define

$$\log(RV_{t-h,h}) \equiv \frac{1}{h}\sum_{i=0}^{h-1}\log(RV_{t-h+i}), \quad \log(RV_{t-1,1}) \equiv \log(RV_{t-1}). \tag{3.2}$$

For example, $\log(RV_{t-22,22})$ averages $\log(RV)$ over the most recent 22 days, that is, from $t-22$ to $t-1$, $\log(RV_{t-5,5})$ over the most recent 5 days, etc.

This leads to our bivariate specification for daily returns and RV with the dynamics of $\log(RV_t)$ modeled as an asymmetric HAR function of past $\log(RV)$. This bivariate system is summarized as follows:

$$r_t = \mu + \epsilon_t,\ \epsilon_t = \sigma_t u_t,\ u_t \sim NID(0,1) \tag{3.3}$$

$$\begin{aligned}\log(RV_t) &= \omega + \phi_1\log(RV_{t-1}) + \phi_2\log(RV_{t-5,5}) + \phi_3\log(RV_{t-22,22}) \\ &+ \gamma u_{t-1} + \eta v_t,\ v_t \sim NID(0,1).\end{aligned} \tag{3.4}$$

This bivariate specification of daily returns and RV imposes the cross-equation restriction that

[6] We assume that any stochastic component in the intraperiod conditional mean is negligible compared to the total conditional variance. It is also straightforward to estimate a bias term.

relates the conditional variance of daily returns to the conditional expectation of daily RV, as in equation (3.1). Joint estimation of the bivariate system in equations (3.3), (3.4) and (3.1) is by maximum likelihood.

Since our applications are to equity returns, it is important to allow for asymmetric effects in volatility. To facilitate comparisons with the benchmark EGARCH model, our parameterization in equation (3.4) includes an asymmetry term, γu_{t-1} associated with the standardized return innovation, u_{t-1}. The impact coefficient for a negative innovation to returns will be $-\gamma$, whereas the impact of a positive innovation will be γ. Typically, $\hat{\gamma} < 0$, which means that a negative innovation to returns implies a higher conditional variance for next period. Unlike EGARCH, our parameterization does not propagate the asymmetry further into future volatility.

In-sample fit of GARCH models have generally favored return innovations with tails that are fatter than those implied by a Normal distribution. Therefore, we evaluate whether or not that result obtains for our bivariate models of returns and RV. That is, we also try replacing equation (3.3) with

$$r_t = \mu + \epsilon_t, \; \epsilon_t = \sigma_t u_t, \;\; u_t \sim t_\nu(0,1), \tag{3.5}$$

in which t_ν denotes a t-distribution with mean 0, variance 1, and ν degrees of freedom. The remainder of the bivariate dynamic system for this case is the same as above. We compare this bivariate system with t-distributed return innovations to that with Normally-distributed innovations, not only for in-sample fit, but also for the term structure of out-of-sample density forecasts.

3.2 Component-RV Specifications

This bivariate specification for daily returns and RV has conditional returns driven by Normal innovations but now the dynamics of $\log(RV_t)$ are captured by two components (2Comp) with different decay rates, as in Maheu and McCurdy (2007). In particular, this bivariate system can be summarized as follows:

$$r_t = \mu + \epsilon_t, \ \epsilon_t = \sigma_t u_t, \ \ u_t \sim NID(0,1) \quad (3.6)$$

$$\log(RV_t) = \omega + \sum_{i=1}^{2} \phi_i s_{i,t} + \gamma u_{t-1} + \eta v_t, \ \ v_t \sim NID(0,1) \quad (3.7)$$

$$s_{i,t} = (1-\alpha_i)\log(RV_{t-1}) + \alpha_i s_{i,t-1}, \ \ 0 < \alpha_i < 1, \ i = 1,2. \quad (3.8)$$

Again, we impose the cross-equation restriction that relates the conditional variance of daily returns to the conditional expectation of daily RV as in equation (3.1). For this specification of our bivariate model, the dynamics of daily $\log(RV)$ are parameterized as the component model specified in equations (3.7) and (3.8) which replace the HAR function in equation (3.4).

Although infinite exponential smoothing provides parsimonious estimates, it possesses several drawbacks. For instance, it does not allow for mean reversion in volatility; and, as Nelson (1990) has shown in the case of squared returns or squared innovations to returns, the model is degenerate in its asymptotic limit. To circumvent these problems, but still retain parsimony, our dynamic model for $\log(RV_t)$, given by equation (3.7), weights each component i by the parameter $0 < \phi_i < 1$ and adds an intercept, ω. Note that when the model is stationary, variance forecasts will mean revert to $\omega/(1-\phi_1-\phi_2)$. This result can be used to do variance targeting and eliminate the parameter ω from the model.[7] This model implies an infinite expansion in $\log(RV_{t-j})$ with coefficients of $\phi_1(1-\alpha_1)\alpha_1^{j-1} + \phi_2(1-\alpha_2)\alpha_2^{j-1}$, $j = 1, 2, ...$[8]

In order to evaluate the potential importance of t-distributed return innovations for this bivariate specification, we replace equation (3.6) with equation (3.5), and jointly estimate with equations (3.7), (3.8) and (3.1).

Motivated by Bollerslev, Kretschmer, Pigorsch, and Tauchen (2009), we also present results for an alternative assumption about available information in which we replace equation (3.1)

[7] That is, set $\omega = \text{mean}(\log(RV))(1-\phi_1-\phi_2)$.

[8] Expanding (3.8) gives $s_{i,t} = (1-\alpha_i)\sum_{n=0}^{\infty} \alpha_i^n \log(RV_{t-1-n})$.

with $\sigma_t^2 \equiv RV_t$. Then

$$r_t = \mu + \epsilon_t,\ \epsilon_t = \sqrt{RV_t}u_t,\ \ u_t \sim NID(0,1) \tag{3.9}$$

$$\log(RV_t) = \omega + \sum_{i=1}^{2} \phi_i s_{i,t} + \gamma u_{t-1} + \eta v_t,\ \ v_t \sim NID(0,1) \tag{3.10}$$

$$s_{i,t} = (1-\alpha_i)\log(RV_{t-1}) + \alpha_i s_{i,t-1},\ \ 0 < \alpha_i < 1,\ i = 1,2. \tag{3.11}$$

which we label 2Comp-OSV.

3.3 Extensions

We consider two extensions to the previous model. The first sets $\eta = 1$, and replaces the innovation v_t in (3.10) with a mixture of two Normals. It has density

$$v_t \sim \begin{cases} N(0,\sigma_{v,1}^2) & \text{with probability } \pi \\ N(0,\sigma_{v,2}^2) & \text{with probability } 1-\pi \end{cases} \tag{3.12}$$

and allows $\log(RV_t)$ to have a fat-tailed distribution.

The second extension is to include GARCH dynamics for the conditional variance of $\log(RV)$. In this case, η in (3.10) has a time subscript and follows the GARCH(1,1) model

$$\eta_t^2 = \kappa_0 + \kappa_1[\log(RV_{t-1}) - E_{t-2}\log(RV_{t-1})]^2 + \kappa_2\eta_{t-1}^2. \tag{3.13}$$

where $\log(RV_{t-1}) - E_{t-2}\log(RV_{t-1})$ denotes the innovation to $\log(RV)$ at time $(t-1)$.

4 Density Forecasts

Our focus is on the return distribution. A popular approach to assess the accuracy of a model's density forecasts is the predictive likelihood (Amisano and Giacomini (2007), Lee, Bao, and Saltoglu (2007), and Weigend and Shi (2000)). This approach evaluates the model's density forecast at the realized return. This is generally done for a one-step-ahead forecast density as multiperiod density forecasts are often not available in closed form. In this paper we advocate multiperiod forecasts since they provide more information to discern among models. The details of the multiperiod predictive likelihood and how to calculate it are described below.

The average predictive likelihood over the out-of-sample observations $t = \tau + k_{max}, ..., T - k$, is

$$D_{M,k} = \frac{1}{T - \tau - k_{max} + 1} \sum_{t=\tau+k_{max}-k}^{T-k} \log f_{M,k}(r_{t+k}|\Phi_t, \theta), \quad k \geq 1, \tag{4.1}$$

where $f_{M,k}(x|\Phi_t, \theta)$ is the k-period ahead predictive density for model M, given Φ_t and parameter θ, evaluated at the realized return $x = r_{t+k}$. Intuitively, models that better account for the data produce larger $D_{M,k}$.

As we will see below for our application to S&P 500, $T = 2936, \tau = 1200$, $k_{max} = 60$ so that $\tau + k_{max} - 1 = 1259$. $D_{M,k}$ is computed for each k using the out-of-sample returns $r_{1260}, ..., r_{2936}$. That is, if $k = 1$, $D_{M,1}$ is computed using out-of-sample returns $r_{1260}, ..., r_{2936}$. For $k = 2$, $D_{M,2}$ is computed using the same out-of-sample returns, etc. This gives us a *term structure of average predictive likelihoods*, $D_{M,1}, ..., D_{M,60}$, to compare the performance of alternative models, M, over an *identical set of out-of-sample data points*.

To assess the statistical differences in $D_{M,k}$ for two models we present Diebold and Mariano (1995) test statistics based on the work of Amisano and Giacomini (2007). Under the null hypothesis of equal performance based on predictive likelihoods of horizen k for models A and B, $t^k_{A,B} = (D_{A,k} - D_{B,k})/(\hat{\sigma}_{AB,k}/\sqrt{T - \tau - k_{max} + 1})$ is asymptotically standard Normal. $\hat{\sigma}_{AB,k}$ is the Newey-West long-run sample variance (HAC) estimate for $d_t = \log f_{A,k}(r_{t+k}|\Phi_t, \hat{\theta}) - \log f_{B,k}(r_{t+k}|\Phi_t, \hat{\theta})$. $\hat{\theta}$ denotes the maximum likelihood estimate for the respective model. Due to the overlapping nature of the density forecasts for $k > 1$ we set the lag-length in the Newey-West variance estimate to the integer part of $[k \times 0.15]$.[9] A large positive (negative) test statistic is a rejection of equal forecast performance and provides evidence in favor of model A (B). As with the predictive likelihoods, a term structure of associated test statistics $t^k_{A,B}$, $k = 1, ..., k_{max}$ are presented in the Results section.

[9] Our results are generally stronger (stronger rejections of the null hypothesis) for smaller lag-lengths.

4.1 Computations

For all $k > 1$ the term $f_{M,k}(r_{t+k}|\Phi_t, \theta)$ will be unknown for the models we consider. However, given that we have fully specified the law of motion for daily returns and RV, we can accurately estimate this quantity by standard Monte Carlo methods. A conventional approach to estimate the forecast density would be to simulate the model out k periods a large number of times and apply a kernel density estimator to these realizations. However, using the kernel density estimator to estimate the forecast density ignores the fact that, in our applications, conditional on the variance we know the distribution. The use of conditional analytic results has been referred to as Rao-Blackwellization and is a standard approach to reduce the variance of a Monte Carlo estimate (Robert and Casella (1999)). This is particularly useful in density estimation which is our context.

To illustrate consider our basic benchmark EGARCH model in (2.3). Note that in this univariate case the information set, Φ_t, just includes past returns. Our estimate is

$$f_{M,k}(r_{t+k}|\Phi_t, \theta) = \int f(r_{t+k}|\mu, \sigma^2_{t+k}) p(\sigma^2_{t+k}|\Phi_t) d\sigma^2_{t+k} \tag{4.2}$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} f(r_{t+k}|\mu, \sigma^{2(i)}_{t+k}), \quad \sigma^{2(i)}_{t+k} \sim p(\sigma^2_{t+k}|\Phi_t) \tag{4.3}$$

where $f(r_{t+k}|\mu, \sigma^{2(i)}_{t+k})$ is a Normal density with mean μ and variance σ^2_{t+k}, evaluated at return r_{t+k}; and $\sigma^{2(i)}_{t+k}$ is simulated out N times according to the EGARCH specification, $p(\sigma^2_{t+k}|\Phi_t)$, which is conditional on time t quantities σ^2_t, u_t, and $\hat{\theta}$, the maximum likelihood estimate of the parameter vector based on Φ_t.

For the joint models of returns and RV, we do a similar exercise to compute the predictive likelihood for returns. In this case, we simulate out both the return and RV dynamics, which implicitly integrates out the unknown σ^2_{t+k}. For each simulation of $RV^{(i)}_{t+1}, ..., RV^{(i)}_{t+k-1}$, $i = 1, ..., N$, we can compute $\sigma^{2(i)}_{t+k} = E_{t+k-1} RV^{(i)}_{t+k}$ using (4.1).[10] A numerical standard error can be used to access accuracy of $\hat{f}_{M,k}(r_{t+k}|\Phi_t, \theta)$ and $\hat{D}_{M,k}$.[11] In our application we found $N = 10000$ to

[10]Recall that the observable SV specification sets $\sigma^{2(i)}_{t+k} = RV^{(i)}_{t+k}$.

[11]To calculate a numerical standard error for $\hat{D}_{M,k}$: let v^2 denote the sample variance of the draws of $f(r_{t+k}|\mu, \sigma^{2(i)}_{t+k})$, then the numerical standard error for $\hat{f}_{M,k}(r_{t+k}|\Phi_t, \theta)$ is $\nu/\sqrt{N}$. Using the delta rule to calculate

provide sufficient accuracy. For example, the numerical standard error is typically well below 1% of $\hat{D}_{M,k}$. Note that for all of our bivariate models the dynamics of the conditional distribution of RV will have a critical impact on the quality of the return density forecasts.

5 Results

Our first results are out-of-sample density forecasts evaluated using predictive likelihoods. The S&P 500 sample starts at 1996/01/02, the first out-of-sample density forecast begins at 2000/12/26 ($t = 1,260$) and ends at 2007/8/29 ($t = 2,936$), for a total of 1,677 density forecasts for each k. We summarize these out-of-sample forecasts by averaging the associated 1,677 predictive likelihoods for each k and then plotting their term structure for the forecast horizons $k = 1, ..., 60$, that is, from 1 to 60 days out of sample. Note that the IBM sample starts at 1993/01/04, the first density forecast begins at 1997/12/24 ($t = 1,260$), and ends at 2007/8/29 ($t = 3,693$), for a total of 2,434 density forecasts for each k. Full sample parameter estimates for the best models are discussed at the end of the section. Model estimation conditions on the first 24 observations.

Our empirical work considered many different models, including different innovation distributions for returns, the value of variance targeting for $\log(RV)$, different functional forms for $\log(RV)$, and a variety of GARCH specifications estimated using daily returns for which EGARCH was the best specification. We note the following general results: models with variance targeting were dominated by the unrestricted version of the model; HAR and component models that link the conditional variance of returns to RV_t by (3.1) always performed better with t-innovations to returns;[12] 2-component models were always better than single-component versions. In the following summary of results, we focus on the top models in different categories.

Our empirical applications to S&P 500 and IBM returns reveal the importance of intraday information, the timing of information availability, and non-Normal innovations to both returns and $\log(RV)$. Figures 2 and 3 compare the term structures of density forecasts for the best models

---

$\hat{Var}(\log \hat{f}_{M,k}(r_{t+k}|\Phi_t, \theta))$; the numerical standard error of $\hat{D}_{M,k}$ is $\sqrt{\sum_{t=\tau+k_{max}-k}^{T-k} \hat{Var}(\log \hat{f}_{M,k}(r_{t+k}|\Phi_t, \theta))}/(T - \tau - k_{max} + 1)$.

[12] We did consider t-innovations for returns in the observable SV models, but estimation supported a Normal distribution since the degree of freedom parameter always moved to extremely large values.