Journal of Econometrics 237 (2023) 105448

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Score-driven models for realized volatility

Andrew Harvey [a], Dario Palumbo [b,c,*]

[a] Faculty of Economics,University of Cambridge, United States of America
[b] Department of Economics, Ca' Foscari University of Venice, Italy
[c] Homerton College, University of Cambridge, United States of America

ARTICLE INFO

Article history:
Received 26 November 2019
Received in revised form 30 November 2022
Accepted 27 January 2023
Available online 22 May 2023

Keywords:
GB2 distribution
HAR model
Heteroscedasticity
Volatility at risk
Weekly volatility pattern

ABSTRACT

This paper sets up a statistical framework for modeling realized volatility (RV) using a Dynamic Conditional Score (DCS) model. It first shows how, for a dataset on stock indices, a preliminary analysis of RV, based on fitting a linear Gaussian model to its logarithm, suggests the use of a two component dynamic specification. It also indicates a departure from normality, a weekly pattern in the data and the presence of heteroscedasticity. Fitting the two component DCS specification with leverage and a day of the week effect is then carried out directly on RV with a Generalized Beta of the Second Kind (GB2) conditional distribution or, equivalently, on the logarithm of RV with an Exponential Generalized Beta of the Second Kind (EGB2) distribution. The forecasting performance of this model, with and without heteroscedasticity, is compared with that of the Heterogeneous Autoregression (HAR), some extensions of it and some other models. Overall there is a clear gain from using the GB2-DCS model, even when the HAR model uses additional information, such as realized semi-variance. When the aim is to forecast tail behavior, the fat-tailed GB2 model performs much better than models with thin-tailed distributions. A further exercise uses a dataset on exchange rates to compare GB2-DCS models with models that use realized quarticity. Again the additional information offers no forecasting advantage. Overall the GB2-DCS models are transparent, provide a comprehensive description of the properties of RV, and are difficult to beat for forecasting.

1. Introduction

This paper sets up a general model for realized volatility (RV) based on a distribution that is coherent for non-negative variables. The dynamics depend on the score of the conditional distribution. Score-driven models were developed in Harvey (2013) and Creal et al. (2013) where they were called dynamic conditional score (DCS) and generalized autoregressive score (GAS) models respectively. A review of recent developments can be found in Harvey (2022). The contribution of this paper is to show how the score-driven approach provides an integrated framework for volatility modeling and how its application to RV yields new insights, as well as giving point forecasts that are competitive with those obtained with existing techniques and are often much better. The flexibility of the conditional distribution also enables them to forecast tail behavior with relative accuracy; the measures used are Volatility-at-Risk, as defined in Caporin et al. (2017), p 134, and the Expected Shortfall for Volatility.

* Correspondence to: Department of Economics, Ca' Foscari University of Venice, Fondamenta S. Giobbe, Cannaregio 873, 30121, Venice, Italy.
E-mail addresses: ach34@cam.ac.uk (A. Harvey), dario.palumbo@unive.it (D. Palumbo).

https://doi.org/10.1016/j.jeconom.2023.01.029

As noted in Opschoor et al. (2018), the intention of realized variance is to measure integrated variance but it may be distorted due to jumps. One way of dealing with this problem is to model RV by a fat tailed distribution. We focus on conditional distributions that are members of the Generalized Beta of the Second Kind (GB2) family. The Burr, Pareto and log-logistic are all special cases of GB2 and the F distribution, used in a multivariate context by Opschoor et al. (2018), is closely related. These distributions have fat tails but the score-driven model handles extreme values robustly. Hence the structure is similar to that of Exponential Generalized Autoregressive Conditional Heteroscedasticity (EGARCH), described in Harvey (2013), chs 4 and 5, and Harvey and Lange (2017). The asymptotic theory in Harvey (2013) is extended to encompass global identifiability and consistency as in the recent paper by Blasques et al. (2022).

Estimating a GB2 for RV is equivalent to estimating a model for its logarithm with an Exponential GB2 (EGB2) distribution. The EGB2 provides further insights into the overall picture because it has the normal distribution as a limiting case. A preliminary analysis of RV can be carried out by fitting a linear Gaussian model to its logarithm. This is easily accomplished using the Kalman filter. In our example, the preliminary analysis confirms the two component specification that has often been found in GARCH and EGARCH models and at the same time reveals a weekly pattern in RV. An analysis of residuals suggests the presence of heteroscedasticity. Fitting the linear Gaussian model also clarifies the relationship with the popular Heterogeneous Autoregression (HAR) model, devised by Corsi (2009) and Müller et al. (1997), which is a simple approximation to the high-order autoregression implied by long memory.

The theoretical background for the GB2 score-driven model is set out in Section 2. Preliminary analysis of the RV from five minute data on four stock indices is described in Section 3 and the results of fitting two component score-driven models with GB2 conditional distributions are reported in Section 4. The preferred models have two first-order components with leverage, together with a day of the week effect. This model is then extended to allow for the heteroscedasticity in the logarithm of RV. Section 5 compares the forecasting performance of DCS models with the performance of modifications of the HAR that use additional information, such as the Continuous-HAR of Andersen et al. (2007) and the Extended HAR of Patton and Sheppard (2015), and two Multiplicative Error Models (MEMs), namely the P-Spline MEM of Brownlees and Gallo (2009) and the MEM-with-Jumps of Caporin et al. (2017). A further exercise, reported in Section 6, compares the DCS models with models that use the measure of realized quarticity (RQ), including those of Bollerslev et al. (2016) and Buccheri and Corsi (2019). The data set used here is for exchange rates because it contained the information needed to construct RQ series, whereas stock index dataset did not. For both data sets forecasts are evaluated according to their ability to predict observations and tail behavior. Section 7 summarizes the results and their implications.

2. Statistical framework

This section sets out a coherent statistical framework for modeling RV that respects the non-negativity of the observations and the tendency of variance to increase with the level; see Corsi et al. (2008) and Taylor (2005), p 335-7. The conditional distribution of the observations is determined by the data.

2.1. Dynamic location/scale models

A dynamic location/scale model for a non-negative variable is

$$y_t = \epsilon_t \mu_{t|t-1}, \quad t = 1, \ldots, T, \tag{1}$$

where the random variable ϵ_t has unit mean. In the MEM model of Engle and Gallo (2006), the conditional mean, $\mu_{t|t-1}$, follows a GARCH-type equation, driven by y_{t-1}, in which the parameters have to be constrained to ensure that y_t remains positive. Estimation is by maximum likelihood (ML), based on the conditional distribution, $f(y_t \mid Y_{t-1})$, with Y_{t-1} denoting the observations up to and including y_{t-1}.

An equivalent model can be formulated in terms of scale, $\alpha_{t|t-1}$, in which case $y_t = \varepsilon_t \alpha_{t|t-1}$, where $E(\varepsilon_t)$, unlike $E(\epsilon_t)$, is not equal to one, unless the distribution is exponential. However, $E(\varepsilon_t)$ is easily found. With an exponential link function for the scale,

$$y_t = \varepsilon_t \exp(\lambda_{t|t-1}), \quad t = 1, \ldots, T, \tag{2}$$

where $\lambda_{t|t-1}$ is unconstrained. First-order dynamics for $\lambda_{t|t-1}$ take the form

$$\lambda_{t+1|t} = \omega(1-\phi) + \phi\, \lambda_{t|t-1} + \kappa\, u_t, \qquad t = 1, \ldots, T, \tag{3}$$

where, in the DCS model, u_t is (proportional to) the score of the conditional distribution of y_t. As with EGARCH, $|\phi| < 1$ ensures stationarity and there is no need to restrict parameters to be non-negative. When the filter, (3), is based on a fat-tailed distribution, the score is bounded. As a result it is not vulnerable to extreme values and it is possible to find restrictions on κ that guarantee invertibility.

Taking logarithms in (2) gives

$$\ln y_t = x_t = \lambda_{t|t-1} + \ln \varepsilon_t, \qquad t = 1, \ldots, T, \tag{4}$$

so $\lambda_{t|t-1}$ now becomes location. A distribution for ε_t implies a distribution for $\ln \varepsilon_t$, so the model can, in principle, be estimated in levels or logarithms.

2.2. GB2 and EGB2 conditional distributions

The usual form of the GB2 density, as given in Kleiber and Kotz (2003), p. 187, is

$$f(y_t;\alpha,\upsilon,\xi,\varsigma)=\frac{\upsilon(y_t/\alpha)^{\upsilon\xi-1}}{\alpha B(\xi,\varsigma)[(y_t/\alpha)^{\upsilon}+1]^{\xi+\varsigma}},\quad 0<y_t<\infty,\quad \alpha,\upsilon,\xi,\varsigma>0, \tag{5}$$

where α is the scale parameter, υ,ξ and ς are shape parameters and $B(\xi,\varsigma)$ is the beta function.

The score with respect to the logarithm of scale is

$$\frac{\partial \ln f_t}{\partial \lambda_{t|t-1}}=u_t=\upsilon(\xi+\varsigma)b_t(\xi,\varsigma)-\upsilon\xi, \tag{6}$$

where

$$b_t(\xi,\varsigma)=\frac{(y_te^{-\lambda_{t|t-1}})^{\upsilon}}{(y_te^{-\lambda_{t|t-1}})^{\upsilon}+1},\qquad t=1,\ldots,T, \tag{7}$$

is distributed as $beta(\xi,\varsigma)$; see Harvey (2013), ch 5. As $y_t\to\infty$, the score approaches an upper bound of $\eta=\upsilon\varsigma$, whereas the lower bound of $-\upsilon\xi$ is obtained by setting $y_t=0$.

Remark 1. The GB2 distribution can be reparameterized so that the (upper) tail index, η, replaces ς, that is we define $\eta=\upsilon\varsigma$. The generalized gamma (GG) is then obtained as a limiting case as $\eta\to\infty$ by replacing α by $\varphi\eta^{1/\upsilon}$. The Weibull is a special case of GG in which $\xi=1$ while setting $\upsilon=1$ gives the gamma distribution. The score for GG is unbounded and for gamma it is linear.

Taking the logarithm of a $GB2(\alpha,\upsilon,\xi,\varsigma)$ gives the exponential GB2 (EGB2) distribution. The EGB2 is symmetric when $\xi=\varsigma$. It includes normal, when $\xi=\varsigma\to\infty$, Laplace, when $\xi=\varsigma\to 0$, and logistic, when $\xi=\varsigma=1$; see McDonald and Xu (1995) and Caivano and Harvey (2014). When y is distributed as $GB2(\alpha,\upsilon,\xi,\varsigma)$ and $x=\ln y$, the PDF of the EGB2 variate x_t is

$$f(x_t;\mu,\upsilon,\xi,\varsigma)=\frac{\upsilon\exp\{\xi(x_t-\lambda)\upsilon\}}{B(\xi,\varsigma)(1+\exp\{(x_t-\lambda)\upsilon\})^{\xi+\varsigma}}. \tag{8}$$

As is evident from (4), the logarithm of scale in GB2 becomes a location parameter in EGB2. Furthermore υ is now a scale parameter, but ξ and ς are still shape parameters and they determine skewness and kurtosis. The distribution has light (exponential) tails. All moments exist with the mean equal to $\lambda+\upsilon^{-1}[\psi(\xi)-\psi(\varsigma)]$ and the standard deviation given by $\sigma=h/\upsilon$, where $h^2=\psi'(\xi)+\psi'(\varsigma)$, with ψ and ψ' denoting the digamma and trigamma functions respectively. The distribution is positively (negatively) skewed when $\xi>\varsigma$ ($\xi<\varsigma$) and its kurtosis decreases as ξ and ς increase; excess kurtosis does not exceed six.

Although υ is a scale parameter in (8), it is the inverse of what would be considered a more conventional measure of scale. There are advantages to parameterizing the model in terms of the standard deviation, σ, which leads to the forcing variable in (3) becoming

$$u_t=\sigma^2\frac{\partial \ln f_t}{\partial \lambda_{t|t-1}}=\sigma h[(\xi+\varsigma)b_t(\xi,\varsigma)-\xi], \tag{9}$$

where $b_t(\xi,\varsigma)$ is as in the GB2, that is (7), but with υ replaced by h/σ and $y_t\exp(-\lambda_{t|t-1})$ replaced by $\exp(x_t-\lambda_{t|t-1})$. Hence the score is bounded as $|x_t|\to\infty$. Parameterizing the model with σ^2 means that letting $\varsigma=\xi\to\infty$ in (9) yields $u_t\to x_t-\lambda_{t|t-1}$; see Caivano and Harvey (2014). The model is then the innovations form of the Kalman filter for a Gaussian state space model consisting of a first-order autoregressive process and an additive error.

An alternative route is to divide the score in (6) by the information quantity for λ which is $\upsilon^2\xi\varsigma/(\xi+\varsigma+1)$. Doing so and replacing υ by h/σ yields

$$u_t=\frac{\xi+\varsigma+1}{h\xi\varsigma}\sigma[(\xi+\varsigma)b_t(\xi,\varsigma)-\xi] \tag{10}$$

and, as with (9), $u_t\to x_t-\lambda_{t|t-1}$ as $\varsigma=\xi\to\infty$.

2.3. Invertibility and asymptotic distribution

A sufficient condition for a filter of the form (3) to be invertible is that $|\phi+\kappa(\partial u_t/\partial\lambda_{t|t-1})|<1$ for all admissible parameters; see Proposition 3.2 (iv) in Blasques et al. (2022). For the GB2 distribution, $x_t=\phi+\kappa u_t'=\phi-\kappa\upsilon^2(\xi+\varsigma)b_t(1-b_t)$. The maximum of $b_t(1-b_t)$ is 1/4 and so the filter must be invertible when $\left|\phi-\kappa\upsilon^2(\xi+\varsigma)/4\right|<1$. Although this constraint is often not restrictive for a GB2, it clearly creates difficulties as ξ and ς become bigger. The problem is avoided when σ replaces υ as the scale parameter and u_t is defined as (9) or (10). In the latter case the invertibility condition becomes

$$\left|\phi-\kappa\frac{(\xi+\varsigma+1)(\xi+\varsigma)}{4\xi\varsigma}\right|<1. \tag{11}$$

When, as usually happens, κ is positive, a sufficient condition for the filter to be invertible is that

$$\kappa < \frac{4(1+\phi)}{\xi+\varsigma}\frac{\xi\varsigma}{\xi+\varsigma+1}. \tag{12}$$

For a logistic distribution, when $\xi = \varsigma = 1$, the invertibility condition is $\kappa < 2(1+\phi)/3$. As $\varsigma = \xi \to \infty$, the distribution becomes normal and a filter based on (10) satisfies the standard invertibility condition for a Gaussian model, namely $|\phi - \kappa| < 1$; see Harvey (2013), p 67.

Fulfillment of the above invertibility conditions guarantees that the (correctly specified) model is globally identifiable and the ML estimators are consistent; see Blasques et al. (2022), Theorem 4.10 and Corollary 4.11. Other conditions, such as the existence of moments, are easily shown to hold and the asymptotic distribution of the ML estimator can be established from Theorem 4.16 of Blasques et al. (2022); see the on-line appendix.

2.4. Dynamic components

A two component model can be used to approximate the features of a long memory process. For RV (and range), the leverage term is governed by $sgn(-r_t)$, where r_t denotes mean-adjusted returns; see Harvey (2013), pp 178–9. Thus

$$\begin{aligned}\lambda_{t|t-1} &= \omega + \lambda_{1,t|t-1} + \lambda_{2,t|t-1},\\ \lambda_{i,t+1|t} &= \phi_i\,\lambda_{i,t|t-1} + \kappa_i\,u_{i,t} + \kappa_i^*\,sgn(-r_t)(u_t+1), \quad i=1,2.\end{aligned} \tag{13}$$

The score curve from the RV model is rotated and, when the distribution of (mean-adjusted) returns is symmetric, the expectation of the composite variable driving the dynamic equation is zero. Identifiability requires $\phi_1 \neq \phi_2$, together with $\kappa_1 \neq 0$ or $\kappa_1^* \neq 0$ and $\kappa_2 \neq 0$ or $\kappa_2^* \neq 0$. An attraction of the two-component model is that the leverage effect can differ in the long and short run.

Other components may be added. Here we employ a day of the week effect so

$$\lambda_{t|t-1} = \omega + \lambda_{1,t|t-1} + \lambda_{2,t|t-1} + \gamma_{t|t-1}, \qquad t = 1,\ldots,T, \tag{14}$$

where $\gamma_{t|t-1}$ has the form of the seasonal component described in Harvey (2013), p 79–80.

2.5. Heteroscedasticity

Corsi et al. (2008) introduce heteroscedasticity, which they call Volatility of Volatility (Vol of Vol), into a model for $\ln RV$ using GARCH combined with a normal inverse Gaussian distribution. The score-driven approach is more easily developed with the EGB2 distribution.

The scale of the EGB2 distribution is $1/\upsilon$. An exponential link function, that is $\upsilon = \exp(-\overline{\upsilon})$, yields

$$\partial \ln f_t/\partial \overline{\upsilon}_{t|t-1} = (\xi+\varsigma)x_t^* b_t - \xi x_t^* - 1, \tag{15}$$

where $x_t^* = (x_t - \lambda_{t|t-1})/\exp(\overline{\upsilon}_{t|t-1})$ and b_t is as in (7) but written in terms of x_t rather than y_t. The score is symmetric but unbounded; see Figure 2 in Caivano and Harvey (2014). Parameterizing in terms of $\sigma_{t|t-1} = h\exp(\overline{\upsilon}_{t|t-1})$ gives an expression for $\partial \ln f_t/\partial \ln \sigma_{t|t-1}$ that is essentially the same as (15).

The (upper) tail index for a GB2 is $\eta = \upsilon\varsigma$ and the lower tail index is $\upsilon\xi$. Thus for fixed values of ς and ξ, a dynamic υ implies that the tail indices are dynamic.

Remark 2. Opschoor and Lucas (2023) model RV measures with an F distribution and extend the model to include dynamic skewness and kurtosis. However, because $\upsilon = 1$ in the F-distribution, heteroscedasticity in the logarithm of RV cannot be modeled. Section 4 discusses further the relationship between the GB2 and F-distributions.

2.6. Long memory, fractional integration and HAR

The standard HAR model, as in Corsi (2009), takes the form

$$\overline{y}_{h,t+h} = \mu + \beta_d y_t + \beta_w \overline{y}_{w,t} + \beta_m \overline{y}_{m,t} + \zeta_{h,t+h}, \qquad h = 1,2,\ldots, \tag{16}$$

where y_t is RV, $\overline{y}_{w,t} = \left(\sum_{i=1}^{5} y_{t-i}/5\right)$, $\overline{y}_{m,t} = \left(\sum_{i=1}^{22} y_{t-i}/22\right)$ and $\overline{y}_{h,t+h}$, $h = 1,2,\ldots$, is the cumulative average over h-days. The disturbance term is $\zeta_{h,t+h}$. In the reparameterization of Patton and Sheppard (2015) the model specification has the second term consisting of only the realized volatilities between lags 2 and 5, and the third term consisting of only the realized volatilities between lag 6 and 22. Usually $h = 1$ in which case $y_{h,t+h} = y_{t+1}$. The coefficients are estimated by regression. The weekly and monthly averages approximate the type of lag structure found with long memory and are seen as a simple alternative to a fractionally integrated model.

As set up in (16), the HAR is a regression model for a variable that is intrinsically positive, with a standard deviation that, in a GB2 framework, is proportional to the location. Working in logarithms, particularly when $h = 1$, stabilizes the variance; see also Corsi et al. (2008) and Corsi and Renò (2012). In this case the $y_t's$ in the summations of (16) are

replaced by their logarithms. The fact that the model for logarithms is additive makes aggregation more appealing because the (conditional) variance is constant.

A recent paper by Baillie et al. (2019) compares the in-sample performance of HAR models with autoregressive fractionally integrated models of the form

$$\Phi(L)(1-L)^d y_t = \varepsilon_t, \quad t = 1, \ldots, T, \tag{17}$$

where the $\Phi(L)$ is a polynomial of order 22 with the same parameter restrictions that are imposed by the HAR. They conclude that the basic HAR fails to capture long memory features adequately.

3. Preliminary analysis

A continuous price process, $P(s)$, is defined by the stochastic differential equation,

$$d\ln P(s) = \mu(s)ds + \sigma(s)dW(s), \qquad 0 \le s \le T, \tag{18}$$

where $\mu(t)$ and $\sigma(t)$ denote the drift and instantaneous volatility processes respectively, and $W(t)$ is a standard Brownian motion. If $r_{t,i}$, $i = 1, 2, \ldots, m$, $t = 1, \ldots, T$, denotes the return in the ith sub-period of day t, the RV measure in day t,

$$RV_t = \sum_{i=1}^{m} r_{t,i}^2, \quad t = 1, \ldots, T, \tag{19}$$

is a consistent estimator of the integrated variance $IV_t \equiv \int_{t-1}^{t} \sigma^2(s)ds$; see, for example, Andersen et al. (2003). The current study is based on high-frequency stock market indices for FTSE 100, DAX, S&P 500 and NASDAQ. The dataset is the one used by Patton and Sheppard (2015) and is constructed from five-minute intraday returns in the Oxford Man Institute library.[1] The observations are for weekdays, with public holidays assumed to be the same as the previous day. The starting date, 3rd January 2000, corresponds to the first non-weekend/holiday date in the U.S. The end date is May 23, 2017. The sample size is therefore $T = 4536$.

When $\ln \varepsilon_t$ in (4) is not too far from normality, fitting a linear state space model to $\ln RV_t$ can provide useful information for model specification. This is easily done using the STAMP package of Koopman et al. (2021). Smoothed estimates of volatility, that is estimates constructed from two-sided filters, are readily available. Taking logarithms also facilitates a comparison with the logarithmic version of HAR.

The preliminary analysis reported here is for FTSE, but investigation of the other indices leads to similar conclusions. The first 23 observations were dropped so as to enable a comparison to be made with HAR. Long memory features are apparent in the ACF of its logarithm. Hence the decision to fit two components. In any case, fitting only one AR(1) component gives a very high portmanteau statistic, $Q(67)$, at 410 and the first-order residual autocorrelation is also high at $r(1) = 0.078$. When two AR(1) components are included, $r(1)$ reduces to 0.033, but Q is still very high and correlations at lags of ten, fifteen and twenty (but not five) are apparent in the residual ACF. Furthermore the associated spectrum shows a clear peak at 2/5 and at its harmonic 4/5. Thus there is strong evidence for a day of the week effect. The model is therefore augmented by including a weekly seasonal component. The seasonals appear to be fixed and the seasonal test statistic, which is asymptotically χ^2_5 under the null of no seasonal effect, is 203.2. The day of the week coefficients are −0.132, −0.052, −0.002, 0.088 and 0.098. Thus Thursday and Friday exhibit much higher volatility than the other days, with Monday being particularly low.[2] The short-term component has a coefficient of 0.75 which, when combined with a long-run coefficient close to unity, implies an ACF similar to that produced by long memory. With the day of the week component included, the implied weights in the filter are as in Fig. 1. There is no evidence of first-order serial correlation as $r(1) = 0.007$ and although the 117.5 value for $Q(70)$ is statistically significant at conventional levels its value is not excessive given the large sample size. The only real failure of the model seems to be in the distribution of the residuals: there is positive skewness and excess kurtosis of about 2.6.

The HAR model does rather well given its simplicity but $Q(67)$ is relatively high at 230.9; the residual autocorrelation is only partly explained by the failure to take account of the day of the week effect. The prediction error variance is 0.240 compared with 0.230 for the unobserved component model. The sum of the coefficients is 0.96 and in the Patton–Sheppard form, the estimates of β_d^*, β_w^* and β_m^* are 0.46, 0.38 and 0.12 respectively; compare the weights in Fig. 1.

4. Fitting score-driven models

Estimating a model with a conditional GB2 distribution with the three parameters υ, ζ and ξ can be problematic. We therefore focus on the Burr distribution, in which $\xi = 1$, and what we call the *balanced* GB2 where $\xi = \varsigma$. The log-logistic distribution is a special case of both Burr and balanced GB2 in which $\varsigma = \xi = 1$, whereas the generalized Pareto distribution is a special case of Burr with $\upsilon = 1$. The logarithm of the balanced GB2 has a symmetric EGB2 distribution

[1] We use the variable 'rv5' from https://realized.oxford-man.ox.ac.uk.

[2] The realized measures in the Oxford library are explicitly constructed so as to ignore any extreme movements in the first few minutes of the trading day, which may be caused by large overnight volatility.

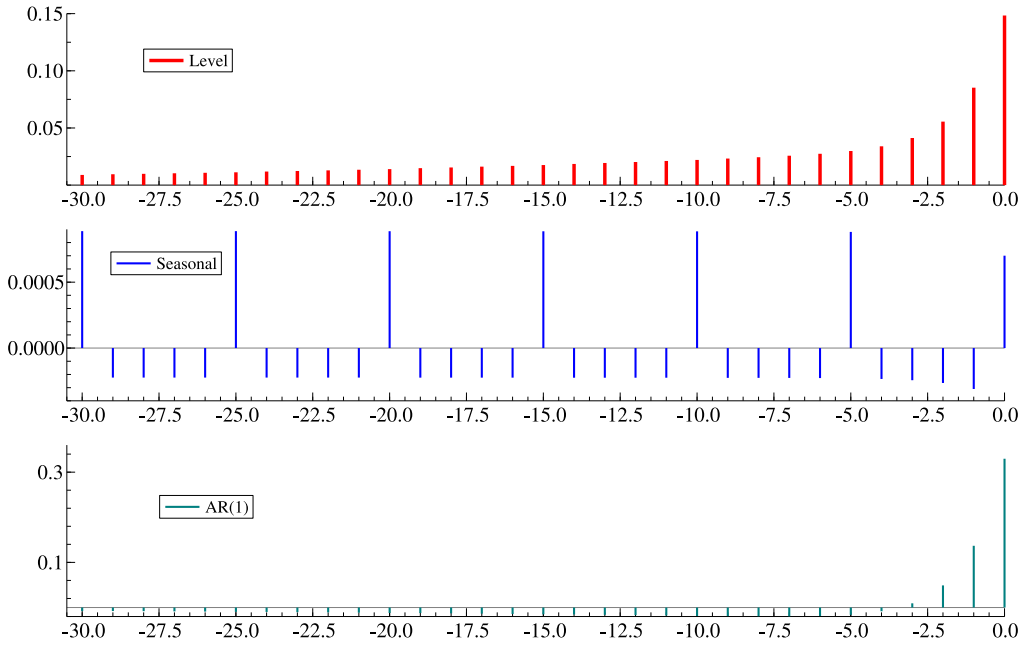

Fig. 1. Filter weights for linear Gaussian model for $\ln FTSE_t$.

which, as noted earlier, goes to a normal distribution as $\xi, \varsigma \to \infty$ meaning that the GB2 becomes lognormal. Finally a variable with an F-distribution with equal degrees of freedom, that is $\nu_1 = \nu_2$, is a special case of a balanced GB2 in which $\upsilon = 1$ and $\xi = \varsigma = \nu_1/2 = \nu_2/2$. More generally when a variable, X, has an F-distribution, $(\nu_1/\nu_2)X$ has a GB2 distribution with $\upsilon = 1$, $\xi = \nu_1/2$ and $\varsigma = \nu_2/2$.

Based on the specification suggested by the preliminary analysis, we fitted these distributions with two components and leverage, as in (13). The Weibull is included for comparison. It is not a fat-tailed distribution, but like the lognormal, it can be obtained from GB2 as a limiting case. Table 1 shows the FTSE results, where Burr is the only distribution not rejected by the Kolmogorov–Smirnov (K–S) test when the significance level is 0.05. The estimate of ς is around 0.8 which implies an EGB2 in which excess kurtosis and skewness are 2.76 and 0.54 respectively; these values are close to those reported for the residuals in the previous section. The results for S&P 500, DAX and NASDAQ are in the on-line appendix. For S&P 500 neither the balanced GB2 and nor the Burr is rejected by the K–S test at the 10% level of significance, but the former gives smaller values for the AIC and BIC. The DAX leans towards Burr while NASDAQ favors balanced GB2. For all four series, the F distribution is not as good as the Burr and balanced GB2 in terms of goodness of fit and it is always rejected by the K–S test.

The values of the tail indices are similar for all four markets, but tend to be lower for the Burr distribution. For example, with FTSE the upper (lower) indices for Burr and balanced F are respectively 3.33 (3.89) and 4.48 (4.48). The estimated υ in balanced GB2 is typically greater than one and the estimates of ξ and ς correspondingly smaller than those reported for F; as a result the tail indices are of a similar magnitude. However, the hypothesis that $\upsilon = 1$, imposed by F, is rejected for all markets by Wald tests on the balanced GB2 estimates. In the log-logistic, where ξ and ς are constrained to be unity, it is υ that adjusts to give comparable tail indices. The fit of the log-logistic is quite good, but the fit with distributions from the generalized gamma class, such as Weibull, is poor.

The low tail indices found in the best GB2 models indicate that the lognormal distribution gives an inferior fit and this is confirmed by the goodness of fit statistics; for FTSE the log-likelihood is 42,067 which is well below the log-likelihood of 42,136 obtained for the balanced GB2 where $\xi = \varsigma = 1.76$ rather than infinity. For S&P, the lognormal log-likelihood is 40,312, again well below balanced GB2. The fit for Pareto is much worse than lognormal and like Weibull it is not considered further.

The day of the week effect shows up clearly in scores from the estimated models, with ACFs similar to those reported for the Gaussian UC model in the previous section. Thus from now on the model labeled DCS has a dynamic equation that includes a seasonal component, as in (14). The results in Table 2 show the parameter estimates for each index with the best-fitting distribution. Taking logarithms and fitting a corresponding EGB2 gives the same estimates and score residuals. The day of the week pattern is similar to that revealed by the preliminary modeling and the leverage effect is stronger in the short-term component, a finding that is not unusual for returns or range; see Harvey and Lange (2018) and Harvey (2013), pp 176–81. Table 3 shows the estimated day of the week effects, together with the estimates of κ_s; these are zero in three cases, indicating that the effect is fixed. For the US indices the most volatile days are Wednesday and Thursday, whereas for the European indices it is Thursday and Friday.

Table 1
Model Selection for the Realized Volatility of FTSE 100 Index. For the Kolmogorov–Smirnov (K–S) statistic the rejections.

Distribution	Shape			Components					Fit			
	ξ	υ	ζ	ω	ϕ_1	κ_1	ϕ_2	κ_2	Logl	AIC	BIC	K–S Test
GB2 $\xi = \zeta$	1.763	2.543	1.763	−8.827	0.999	0.036	0.842	0.058	42,135.76	- 84,257.52	- 84,212.58	0.027
	(0.136)	(0.085)	(0.136)	(0.415)	(0.001)	(0.007)	(0.040)	(0.006)				(0.002)
Burr	1	3.893	0.856	−8.796	0.999	0.033	0.867	0.056	42,130.52	- 84,247.03	- 84,202.09	0.020
	–	(0.022)	(0.048)	(0.383)	(0.001)	(0.006)	(0.030)	(0.006)				(0.052)
Log-Logistic	1	3.676	1	−8.764	0.999	0.034	0.862	0.056	42,125.46	- 84,238.92	- 84,200.40	0.023
	–	(0.012)	–	(0.387)	(0.001)	(0.006)	(0.033)	(0.006)				(0.020)
F	3.275	–	2.627	−8.883	0.826	0.062	0.999	0.037	42,119.52	- 84,225.03	- 84,180.09	0.037
	(0.075)	–	(0.042)	(0.430)	(0.042)	(0.006)	(0.001)	(0.006)				(0.000)
F $nu_1 = nu_2$	4.494	1	4.494	−8.919	0.999	0.039	0.814	0.062	42,099.74	- 84,187.48	- 84,148.96	0.041
	(0.020)	–	(0.020)	(0.469)	(0.002)	(0.007)	(0.049)	(0.007)				(0.000)
Pareto	1	1	39.499	−9.789	0.991	0.141	0.580	0.207	40,012.12	- 80,012.24	- 79,973.72	0.284
	–	–	(0.187)	(0.236)	(0.004)	(0.030)	(0.174)	(0.033)				(0.000)
Weibull	1	1.330	∞	−9.688	0.992	0.023	0.667	0.131	40,433.65	- 80,855.29	- 80,816.77	0.167
	–	(0.008)	–	(0.073)	(0.001)	(0.003)	(0.050)	(0.011)				(0.000)

* denotes a 10% significance level, ** a 5% significance level and *** a 1% significance level.

Table 2
Estimation results from DCS and DCS-H models with a daily component and leverage.

Index	Shape			Components							Heteroscedasticity		
	ξ	υ	ζ	ω_λ	$\phi_{\lambda 1}$	$\kappa_{\lambda 1}$	$\kappa_{\lambda 1l}$	$\phi_{\lambda 2}$	$\kappa_{\lambda 2}$	$\kappa_{\lambda 2l}$	ω_υ	ϕ_υ	κ_υ
FTSE 100	1	4.034	0.825	−8.699	1.000	0.029	0.000	0.887	0.052	0.018			
	–	(0.022)	(0.047)	(0.386)	(0.001)	(0.004)	(0.003)	(0.019)	(0.005)	(0.004)			
	1		0.817	−8.620	1.000	0.028	0.000	0.892	0.049	0.018	−1.400	0.510	0.045
	–		(0.047)	(0.366)	(0.001)	(0.004)	(0.003)	(0.017)	(0.005)	(0.004)	(0.023)	(0.139)	(0.009)
DAX	1	3.800	0.851	−7.792	1.000	0.036	0.004	0.846	0.056	0.025			
	–	(0.022)	(0.049)	(0.523)	(0.001)	(0.005)	(0.004)	(0.028)	(0.005)	(0.004)			
	1		0.851	−7.810	1.000	0.035	0.004	0.857	0.054	0.023	−1.322	0.985	0.008
	–		(0.049)	(0.473)	(0.001)	(0.004)	(0.003)	(0.026)	(0.005)	(0.004)	(0.029)	(0.009)	(0.003)
S&P 500	2.017	1.947	2.017	−8.408	1.000	0.039	−0.001	0.866	0.090	0.052			
	(0.153)	(0.094)	(0.153)	(0.723)	(0.002)	(0.006)	(0.006)	(0.018)	(0.007)	(0.007)			
	2.161		2.161	−8.703	1.000	0.039	0.001	0.860	0.085	0.049	−0.627	0.531	0.037
	(0.157)		(0.157)	(0.813)	(0.003)	(0.005)	(0.006)	(0.020)	(0.006)	(0.004)	(0.095)	(0.132)	(0.008)
Nasdaq 100	2.443	2.032	2.443	−7.716	1.000	0.029	0.001	0.837	0.089	0.026			
	(0.173)	(0.103)	(0.173)	(0.378)	(0.020)	(0.006)	(0.004)	(0.001)	(0.005)	(0.003)			
	2.743		2.743	−8.139	1.000	0.030	0.001	0.820	0.081	0.026	−0.646	0.556	0.044
	(0.175)		(0.175)	(0.367)	(0.001)	(0.005)	(0.003)	(0.024)	(0.005)	(0.004)	(0.102)	(0.092)	(0.007)

Table 3
Estimation results from DCS model with a daily component and leverage - daily component.

Index	Seasonal					
	κ_s	γ_1	γ_2	γ_3	γ_4	γ_5
FTSE 100	0.001	−0.119	−0.029	0.031	0.043	0.075
	(0.001)	(0.020)	(0.015)	(0.021)	(0.026)	–
DAX	0.000	−0.113	−0.044	0.004	0.081	0.072
	(0.000)	(0.007)	(0.016)	(0.012)	(0.007)	–
S&P	0.000	−0.129	−0.006	0.063	0.069	0.003
	(0.002)	(0.075)	(0.015)	(0.014)	(0.068)	–
Nasdaq	0.001	−0.114	−0.045	0.122	0.045	−0.007
	(0.000)	(0.031)	(0.034)	(0.028)	(0.028)	–

Although the residuals from the preferred Gaussian model have relatively little serial correlation, this is not the case for their squares or absolute values. The $Q-$statistics formed from the scores for υ, or equivalently σ, tell a similar story. This suggests the need to model heteroscedasticity in $\ln RV_t$. Estimation of the GB2/EGB2 model with two components, as in (13), seasonal dummies and a dynamic υ, captured by a first-order equation for $\overline{\upsilon}_{t|t-1}$, gives the second set of results shown for each index in Table 2. In all cases apart from DAX, the dynamic effects in $\upsilon_{t|t-1}$ are short-lived with the autoregressive coefficient only slightly bigger than 0.5. Nevertheless, the information criteria in Table 4 indicate that their inclusion in what we call the DCS-H model improves the fit. Furthermore there is a big reduction in the value of the portmanteau statistics for the scores of υ, although not by enough to completely eliminate the residual heteroscedasticity.

Table 4
Goodness of fit comparison for models with and without heteroscedasticity. The Q tests are computed on the fitted scores with respect to υ.

Index	Model	Logl	AIC	BIC	$Q_\upsilon(5)$	$Q_\upsilon(20)$
FTSE 100	DCS	42,236.70	− 84,447.39	− 84,363.93	122.454***	157.988***
	DCS-H	42,270.43	− 84,508.86	− 84,406.14	53.705***	84.469***
DAX	DCS	38,727.42	− 77,426.83	− 77,336.95	131.337***	209.129***
	DCS-H	38,748.30	− 77,464.60	− 77,361.88	73.715***	95.681***
S&P 500	DCS	40,466.07	− 80,904.14	− 80,814.26	112.195***	188.159***
	DCS-H	40,481.79	− 80,931.59	− 80,828.87	22.111***	78.503***
Nasdaq 100	DCS	40,505.39	− 80,982.79	− 80,892.91	157.272***	246.286***
	DCS-H	40,539.83	− 81,047.66	− 80,944.94	24.459***	86.623***

*** denotes a nominal 1% significance level (No adjustment for fitted parameters).

5. Forecasting

In order to compare the forecasting performance of the DCS and DCS-H models with the benchmark HAR models on the stock market datasets used in the last section, we identified a nine year forecasting period from 01/01/2008 to 30/12/2016. The time span is long enough to include a number of major financial market events and crises. Models were estimated using an eight year moving window up to each out-of-sample date and one-step-ahead forecasts were computed.

The basic HAR models are the one in levels, as formulated by Patton and Sheppard (2015), the model of Corsi (2009) in logarithms, and the levels RARFIMA of Baillie et al. (2019), as defined in (17). The one step ahead forecasts for the mean of the GB2 model are obtained by making a modification to the forecast of the scale; see Harvey (2013), p 163. In the general case, with heteroscedasticity,

$$\widetilde{y}_{t|t-1} = \frac{\Gamma\left(\xi + 1/\upsilon_{t|t-1}\right)\Gamma\left(\varsigma + 1/\upsilon_{t|t-1}\right)}{\Gamma(\xi)\,\Gamma(\varsigma)} e^{\lambda_{t|t-1}}, \quad t = T+1, \ldots, T+T^*,$$

where T^* denotes the number of observations in the forecasting period. When the HAR model is estimated in logarithms, the standard lognormal correction factor is used to give $\widehat{y}_{t|t-1} = \exp\left(\widehat{x}_{t|t-1} + \hat{\sigma}^2/2\right)$.

5.1. Extensions to HAR

The comparison was extended to include some HAR models that use additional information extracted from raw returns. These are the Continuous-HAR (CHAR), proposed by Andersen et al. (2007) and the Extended HAR (EHAR) in Patton and Sheppard (2015). The first model is based on the idea that the total variation of RV_t can be decomposed into a continuous and a discontinuous component. Following Barndorff-Nielsen and Shephard (2004), this decomposition can be implemented through the Bi-Power Variation (BPV), defined as

$$BPV_t = z_t = \frac{\pi}{2}\sum_{i=1}^{m-1}\left|r_{t_i}\right|\left|r_{t_i-1}\right|, \qquad t = 1, 2, \ldots, T.$$

The series on BPV_t for the four indices come from the same Oxford Man Realized Library dataset. The CHAR model replaces RV by BPV on the right hand side of the levels HAR specification, so that

$$\bar{y}_{h,t+h} = \mu + \beta_d z_t + \beta_w \bar{z}_{w,t} + \beta_m \bar{z}_{m,t} + \zeta_{h,t+h}, \quad h = 1, 2, \ldots \tag{20}$$

In the EHAR model, the contemporaneous realized variance, y_t, on the right hand side of (16) is replaced by realized semi-variances, which are realized variances computed using only returns with concordant signs, that is

$$RSV_t^+ = \sum_{i=1}^{m} r_{t_i}^2 I_{\{r_{t_i}>0\}} \quad \textit{and} \qquad RSV_t^- = \sum_{i=1}^{m} r_{t_i}^2 I_{\{r_{t_i}\le\}}.$$

The specification of the EHAR is therefore

$$\bar{y}_{h,t+h} = \mu + \beta_d^+ RSV_t^+ + \beta_d^- RSV_t^- + \beta_w \bar{y}_{w,t} + \beta_m \bar{y}_{m,t} + \zeta_{h,t+h}, \qquad h = 1, 2, ..$$

We also included some popular MEMs, namely the P-Spline MEM of Brownlees and Gallo (2009), which uses a spline function for the mean,[3] and the MEM-with-Jumps (MEM-J) of Caporin et al. (2017), which is based on a mixture of gamma type distributions.[4] The MEM-J distribution involves an infinite sum, which needs to be truncated for tractability. Overall it is more computationally intensive than a GB2, and much less straightforward to work with. Furthermore, although the mixture distribution may accommodate jumps, the dynamic equation for the conditional mean is driven by y_{t-1} and so, unlike the GB2-DCS model, it will not be resistant to outliers.

[3] We used 16 knots equally spaced knots as this provided the best fit in-sample.

[4] We used the specification which provided the best in-sample fit. This had four jump components.

Table 5
The table summarizes the relative forecasting performance of each of the twelve models considered on the indexes dataset in terms of the accuracy measures of RMSFE, MAFE and Qlike. Rows marked "I" report the number of times a model outperforms other models, while rows marked "II" report the number of times the model is outperformed. For each comparison, the number of times the DM statistic is significant at the 10% level is shown in brackets.

			DCS	DCS-H	HAR	Log HAR	CHAR	EHAR	RARFIMA	Spline-MEM	MEM-J
RMSFE	FTSE	I	3(0)	2(0)	4(0)	6(1)	–	7(2)	1(0)	5(0)	8(1)
		II	5(0)	6(0)	4(1)	2(0)	8(2)	1(0)	7(1)	3(0)	–
	DAX	I	7(1)	2(1)	1(0)	3(0)	6(0)	4(0)	5(0)	–	8(2)
		II	1(0)	6(0)	7(1)	5(0)	2(0)	4(1)	3(0)	8(0)	–
	S&P	I	8(1)	7(1)	1(1)	4(1)	6(2)	2(1)	3(1)	–	5(1)
		II	–	1(0)	7(1)	4(0)	2(0)	6(0)	5(0)	8(8)	3(0)
	Nasdaq	I	5(0)	2(0)	–	6(1)	4(0)	7(3)	1(0)	3(0)	8(5)
		II	3(0)	6(0)	8(2)	2(1)	4(2)	1(0)	7(2)	5(2)	–
	Total	I	23(2)	13(2)	6(1)	19(3)	16(2)	20(6)	10(1)	8(0)	29(9)
		II	9(0)	19(0)	26(5)	13(1)	16(4)	12(1)	22(3)	24(10)	3(0)
MAFE	FTSE	I	7(2)	8(2)	4(3)	3(2)	–	5(4)	1(1)	2(1)	6(3)
		II	1(0)	–	4(1)	5(0)	8(8)	3(0)	7(6)	6(3)	2(0)
	DAX	I	8(6)	6(5)	4(1)	3(1)	1(1)	5(4)	2(1)	–	7(6)
		II	–	2(0)	4(3)	5(3)	7(4)	3(1)	6(4)	8(8)	1(0)
	S&P	I	7(6)	6(6)	1(0)	4(1)	8(6)	3(0)	2(1)	–	5(0)
		II	1(0)	2(0)	7(3)	4(3)	–	5(3)	6(3)	8(5)	3(3)
	Nasdaq	I	7(5)	8(5)	3(3)	4(3)	2(0)	5(3)	–	1(0)	6(6)
		II	1(0)	–	5(3)	4(3)	6(6)	3(1)	8(6)	7(6)	2(0)
	Total	I	29(19)	28(18)	12(7)	14(7)	11(7)	18(11)	5(3)	3(1)	24(15)
		II	3(0)	4(0)	20(10)	18(9)	21(18)	14(5)	27(19)	29(22)	8(3)
Qlike	FTSE	I	2(1)	1(1)	7(4)	3(1)	4(1)	8(8)	–	6(1)	5(1)
		II	6(1)	7(2)	1(1)	5(2)	4(2)	–	8(8)	2(1)	3(1)
	DAX	I	4(0)	3(0)	5(2)	2(1)	6(3)	7(4)	–	1(1)	8(4)
		II	4(0)	5(0)	3(2)	6(3)	2(0)	1(0)	8(6)	7(4)	–
	S&P	I	6(3)	5(3)	2(2)	3(2)	7(4)	4(3)	1(1)	–	8(4)
		II	2(0)	3(0)	6(5)	5(1)	1(0)	4(1)	7(7)	8(8)	–
	Nasdaq	I	5(1)	3(1)	7(2)	4(1)	1(1)	6(2)	–	2(1)	8(7)
		II	3(0)	5(1)	1(1)	4(1)	7(3)	2(1)	8(8)	6(1)	–
	Total	I	17(5)	12(5)	21(10)	12(5)	18(9)	25(17)	1(1)	9(3)	29(16)
		II	15(1)	20(3)	11(9)	20(7)	14(5)	7(2)	31(29)	23(14)	3(1)

5.2. Forecast comparison

The ability of the models to forecast RV_t is compared using root mean squared forecast errors (RMSFE), mean absolute forecast errors (MAFE) and the QLike measure

$$QLike(y_{t|t-1}) = y_{t|t-1}/\widehat{y}_{t|t-1} - \ln\left(y_{t|t-1}/\widehat{y}_{t|t-1}\right) - 1, \quad t = T+1, \ldots, T+T^*.$$

Patton (2011) shows that only RMSFE and QLike yield a coherent ranking of forecasts of y_t, but because MAFE is widely used we include it as well. Having computed these measures for all models, we carried out the Diebold and Mariano (2002) test, denoted DM, for all pairs to assess whether the differences were significant. Table 5 provides a comprehensive summary of the results, reporting how many times each model outperforms, or is outperformed, by others in each of the measures as well as, in parenthesis, how many times the outperformance is significant with respect to the DM test results.[5] Before commenting on these results, we first note that Table 6 shows that the preferred DCS models beat all the others in terms of predictive likelihood; see Mitchell and Wallis (2011).

Overall it is difficult to identify a model that outperforms all the others and there is often a discrepancy between the three measures. According to the RMSFE and MAFE measures, the HAR performs better when formulated in logarithms, while the HAR in levels, as well as the CHAR and EHAR, are preferred by QLike. The worst performer is the RARFIMA followed by the Spline-MEM.[6] Perhaps RARFIMA models for the logarithm of RV would fare better.

The best performer according to the RMSFE measure is the MEM-J, while according to MAFE it is the DCS and according to QLike it is the EHAR. In all the cases the MEM-J, despite not always providing as many significant results, tends to be among the best models despite being significantly outperformed on a number of occasions. By contrast, only once is the

[5] We also provide in the online appendix individual detailed tables with the results of the tests and the magnitude of the relative performance of the models in terms of ratio of loss functions.

[6] Harvey and Lange (2017) found that spline GARCH models performed poorly relative to Beta-t-EGARCH.

Table 6
Results of UC and UB tests for FTSE, DAX, S&P and Nasdaq data series.

		p-quantiles	DCS	DCS-H	HAR	HAR in Logs	CHAR	EHAR	RARFIMA	Spline-MEM	MEM-J
		0.1	0.072	0.378	297.099***	62.386***	332.679***	297.099***	305.569***	19.764***	5.587**
	UC	0.05	0.272	0.001	121.226***	35.766***	135.744***	128.283***	121.226***	2.540	20.325***
		0.01	0.011	0.276	3.555*	0.097	5.644**	2.716*	4.528**	6.680***	22.828***
FTSE		0.1	0.390	−0.063	−11.17***	−6.466***	−11.71***	−11.23***	−13.14***	−2.442**	4.963***
	UB	0.05	0.698	0.159	−6.824***	−4.269***	−7.140***	−6.830***	−9.196***	0.193	6.586***
		0.01	1.135	0.629	0.551	0.525	−0.125	0.858	−4.213***	5.883***	5.611***
	Pred Lik		21,726.39	21,735.69	17,980.57	21,551.59	17,923.71	18,015.76	17,846.97	21,437.23	21,614.55
		0.1	0.924	1.354	258.283***	41.599***	273.154***	261.925***	265.617***	15.970***	7.251***
	UC	0.05	4.304**	2.941*	88.268***	20.348***	93.621***	90.913***	90.913***	1.213	7.755***
		0.01	5.787**	4.173**	0.276	0.096	0.276	0.276	1.409	19.843***	2.797*
DAX		0.1	1.918*	1.329	−10.12***	−4.723***	−10.40***	−10.27***	−12.32***	−1.472	3.177***
	UB	0.05	2.071**	1.592	−5.427***	−2.632***	−5.725***	−5.456***	−8.309***	1.661*	2.757***
		0.01	3.630***	2.680***	2.438**	2.084**	2.287**	2.397**	−3.174***	8.986***	3.936***
	Pred Lik		20,031.36	20,031.51	16,470.52	19,876.09	16,564.27	16,510.20	16,536.08	19,712.63	19,960.91
		0.1	11.633***	7.251***	162.792***	10.135***	148.540***	167.762***	93.431***	252.899***	14.431***
	UC	0.05	25.124***	15.311***	37.177***	6.251**	31.735***	41.622***	8.479***	368.710***	24.292***
		0.01	18.414***	15.687***	5.787**	1.675	10.785***	9.681***	38.198***	409.946***	18.414***
S&P		0.1	5.870***	4.567***	−7.194***	−2.140**	−6.856***	−7.296***	−10.76***	26.678***	5.572***
	UB	0.05	6.969***	5.260***	−2.095**	−1.073	−1.882*	−2.050**	−6.427***	32.381***	5.925***
		0.01	6.824***	6.059***	7.550***	2.655***	8.670***	8.421***	−0.027	44.705***	6.495***
	Pred Lik		20,764.84	20,775.81	14,933.37	20,677.43	15,145.20	14,941.33	15,897.23	19,505.36	20,566.09
		0.1	1.864	1.202	314.309***	29.547***	297.099***	284.865***	327.968***	2.787*	23.872***
	UC	0.05	9.289***	8.763***	128.283***	18.423***	121.226***	111.321***	131.961***	0.640	21.093***
		0.01	13.141***	4.951**	2.003	0.097	2.003	1.409	2.003	21.314***	1.675
Nasdaq		0.1	2.719***	2.553**	−11.31***	−4.568***	−11.14***	−10.86***	−12.98***	1.045	5.446***
	UB	0.05	3.628***	3.094***	−6.589***	−3.193***	−6.488***	−6.154***	−8.668***	3.587***	3.815***
		0.01	4.904***	3.752***	1.601	0.532	1.544	1.689*	−3.250***	7.356***	1.739*
	Pred Lik		21,338.98	21,345.31	17,421.06	21,265.94	17,443.76	17,595.22	17,308.11	21,095.46	21,243.06

* denotes a 10% significance level, ** a 5% significance level and *** a 1% significance level. We also report the predictive likelihood for each model.

DCS outperformed significantly. Thus, despite not always providing the best model, the DCS models are never significantly outperformed, something that is not true for the other models.

5.3. Density and tail forecasting

As Corsi et al. (2008) point out, producing accurate volatility forecasts is not only a matter of producing accurate point forecasts. Misspecified models can yield relatively accurate point forecasts while grossly underestimating or overestimating the tail size and hence the probability of occurrence of extreme events. In order to investigate this we compared the ability of each of the models to produce accurate out-of-sample density forecasts.

We follow Caporin et al. (2017), p 134, and define the one-step-ahead p-upper Volatility-at-Risk (VolaR) as the predicted $1-p$ quantile, that is $VolaR_{t|t-1}(p) = F^{-1}_{t|t-1}(1-p)$, where $F^{-1}_{t|t-1}(1-p)$ is the conditional quantile function. We also consider the one-step-ahead upper expected shortfall, called Expected Shortfall for Volatility, that is

$$ESVol_{t|t-1}(p) = E\left[y_t | Y_{t-1} \geq VolaR_{t|t-1}(p)\right] = \frac{1}{p}\int_{VolaR_{t|t-1}(p)}^{\infty} x f_{t|t-1}(x)\, dx,$$

for $0 < p < 1$ and $t = T+1, \ldots, T+T^*$.

By defining the hit-process $h_t(p) = 1$ when $y_t > VolaR_{t|t-1}(p)$ and zero otherwise, we can construct the unconditional coverage (UC) likelihood ratio tests of Christoffersen (1998) for the VolaR violations. The null hypothesis is that $E[h_t(p)] = p$ and the test statistic is asymptotically $\chi^2(1)$. To evaluate the accuracy in forecasting the ESVol we use the unconditional backtest (UB) of Du and Escanciano (2017). The UB test statistic is constructed from the cumulative violation process, $H_t(p) = p^{-1}(p - (1 - PIT_t))\, h_t(p)$, where PIT_t denotes the conditional one-step-ahead probability integral transform, $F_{t|t-1}\left(\hat{y}_{t|t-1}\right)$. The '$t$ statistic', which is asymptotically standard normal under the null hypothesis that $E[H_t(p)] = p/2$, is

$$t_{ESVol} = \frac{\overline{H}(p) - p/2}{\sqrt{p\,(1/3 - p/4)\,/T^*}},$$

where $\overline{H}(p)$ denotes the sample mean of the $H_t(p)'s$.

Table 6 shows the results of the two tests for $p = 0.10$, 0.05 and 0.01, together with the predictive likelihood. For the European indices, the DCS model with a Burr distribution gave the best fit. In this case the quantile function takes a particularly simple form so that $VolaR_{t|t-1}(p) = 1 - \exp\left(\lambda_{t|t-1}\right)\left(p^{-1/\varsigma} - 1\right)^{1/\upsilon_{t|t-1}}$. More generally, VolaR and ESVol can be calculated with routines for the evaluation of incomplete beta functions and their inverses; see Kleiber and Kotz (2003), p 188, 192.

Comparing the results of the UC and UB tests we see that the HAR models tend to seriously overestimate the quantiles for $p = 0.1$ and 0.05. There is an improvement for $p = 0.01$, but the tests still tend to reject. The HAR in logarithms mitigates the magnitude of the rejections, but the picture is still the same. The consequences could be an unreasonable hedge in a portfolio to protect against volatility risk and a tendency to overestimate a volatility spillover effect in a systemic risk assessment. This point should be considered carefully by risk managers and policy makers; see also Lucas et al. (2014) where implied default probabilities in Credit Default Swap prices are modeled. The magnitudes of the test statistics for the MEM models tend to be smaller than for the HAR models except for $p = 0.01$. However, the DCS models

have the smallest number of rejections across all quantiles, with DCS-H being the only model to show no rejections for the FTSE dataset.

The results for the American indices show a larger number of rejections. The DCS models, now driven by balanced GB2 conditional distributions, are still better than the HAR models. The performance of the MEM-J model is comparable to that of the DCS models, but that of Spline-MEM is much worse for S&P. However, it should be borne in mind that the MEM-J model takes much longer to estimate than a DCS model.

6. Forecasting using realized quarticity

Recent developments in the literature have shown that RV_t is a noisy estimator of the IV of the true continuous price process. Following the asymptotic theory of Bandorff-Nielsen and Shephard (2002), RV_t can be decomposed as

$$RV_t = IV_t + \eta_t \qquad \eta_t \sim MN\left(0, (2/m)IQ_t\right), \qquad t = 1, \ldots, T, \tag{21}$$

where $IQ_t \equiv \int_{t-1}^{t} \sigma_s^4 ds$ is called Integrated Quarticity, m is the number of intraday returns used to construct RV_t and MN stands for mixed normal, that is a normal distribution conditional on IQ_t. This result implies deviations from the asymptotic normality of RV_t. To improve the performance of the HAR model, Bollerslev et al. (2016) introduced the measure of Realized Quarticity,

$$RQ_t = \frac{m}{3}\sum_{i=1}^{m} r_{t_i}^4, \qquad t = 1, \ldots, T,$$

as a consistent estimator of IQ_t. They showed that the impact of the variation of IQ_t over time can be captured by time varying parameters in the HAR model dependent on IQ_t. Hence they replaced β_d in (16) by $\beta_d + \beta_Q RQ_t^{1/2}$ to give the HARQ model. They also proposed the full HARQ (HARQF) model

$$\bar{y}_{h,t+h} = \mu + \left(\beta_d + \beta_{dQ} RQ_t^{1/2}\right) y_t + \left(\beta_{1w} + \beta_{wQ}\overline{RQ}_{w,t}^{1/2}\right)\bar{y}_{w,t} + \left(\beta_{1m} + \beta_{mQ}\overline{RQ}_{m,t}^{1/2}\right)\bar{y}_{m,t} + \zeta_{h,t+h}.$$

As an alternative, Buccheri and Corsi (2019) proposed modeling the relation between RV_t and IV_t in logarithms, so that

$$\ln RV_t = \ln IV_t + \xi_t \qquad \xi_t \sim MN\left(0, \frac{2}{m}\frac{IQ_t}{IV_t^2}\right). \tag{22}$$

The IQ_t and IV_t variables are replaced by their estimators, RQ_t and RV_t, in the resulting "SHARK" model, the unobserved signal, $\ln IV_t$, evolves as a HAR with time varying parameters and models dynamically heteroscedasticity. Estimation is by the Kalman filter, but the large state vector means that the computations can be time-consuming.

Does the inclusion of RQ_t offer significant improvement in the forecasting of RV_t? Our stock index source did not contain information on IQ, so in order to make an assessment we performed a forecasting comparison on three RV series of foreign exchange, namely EUR/USD, EUR/GBP and GBP/USD, constructed from five minute intraday returns from Monday to Friday between 04/03/2002 and 31/12/2013.[7] From this we constructed our daily volatility measures with roughly 3000 observations. We then split the dataset in two parts, leaving an out-of-sample period of six years, from 01/01/2008 to 31/12/2013. We give results for all the models considered so far,[8] as well as for the HARQ, HARQF and the SHARK models. All the models are re-estimated at each time step assuming as estimation period a moving window of six years. Forecasting accuracy of the models is then compared using the same point and density criteria as in the previous subsections.

In all the three datasets the DCS models have the largest predictive likelihoods, but the conclusions are less clear cut for other measures. We can see in Table 7 that HARQ outperforms significantly the largest number of models according to the RMSFE and MAFE measures, while performing worse than DCS-H according to the QLike measure. Under this measure the best performer is the CHAR model. Across all the measures, the HAR in logarithms performs significantly worse than the HAR as well as the Spline-MEM and MEM-J. The SHARK performs well according to the RMSFE but is outperformed by the DCS models according to MAFE and QLike. The worst performer remains the RARFIMA. However, as with the previous datasets, the DCS models are outperformed significantly by the smallest number of models after the CHAR. It seems that the DCS models are unlikely to yield a bad forecasting performance irrespective of the datasets and forecasting period.

As regards density forecasting, Table 8 shows that the DCS models again tend to have the smallest UC and UB statistics across all the quantiles, together with the smallest number of rejections. The MEM models capture the distribution better

[7] We obtained the high frequency dataset from http://www.histdata.com/. Most of the series there seem to be incomplete so we have used the ones that had less than 1% of missing data once re-sampled at the 5 min frequency. The datasets we constructed are available upon request.

[8] The in-sample analysis revealed the same features of weekly seasonality as in the stock market indexes. The balanced GB2 distribution gave the best fit. We used the same dynamic components specification as in the previous section but without leverage.

Table 7
The table summarizes the relative forecasting performance of each of the twelve models considered on the exchange rates dataset in terms of the accuracy measures of RMSFE, MAFE and Qlike. Rows marked "I" report the number of times a model outperforms other models, while rows marked "II" report the number of times the model is outperformed. For each comparison, the number of times the DM statistic is significant at the 10% level is shown in brackets.

			DCS	DCS-H	HAR	Log HAR	CHAR	EHAR	HARQ	HARQF	SHARK	RARFIMA	Spline-MEM	MEM-J
	EUR/USD	I	6(1)	5(1)	10(2)	4(1)	11(1)	9(2)	8(4)	7(1)	1(1)	–	3(1)	2(1)
		II	5(0)	6(0)	1(0)	7(1)	–	2(0)	3(2)	4(0)	10(0)	11(11)	8(1)	9(1)
	EUR/GBP	I	7(6)	8(6)	5(4)	3(3)	9(5)	7(6)	11(11)	1(1)	10(10)	–	2(2)	4(4)
		II	4(3)	3(2)	6(5)	8(6)	2(2)	4(2)	–	10(10)	1(1)	11(11)	9(9)	7(7)
RMSFE	GBP/USD	I	7(4)	5(4)	10(7)	4(4)	9(5)	11(5)	3(2)	2(2)	1(1)	–	8(2)	6(2)
		II	4(2)	6(2)	1(0)	7(3)	2(2)	–	8(5)	9(5)	10(8)	11(11)	3(0)	5(0)
	Total	I	20(11)	18(11)	25(13)	11(8)	29(11)	27(13)	22(17)	10(4)	12(12)	–	13(5)	12(7)
		II	13(5)	15(4)	8(5)	22(10)	4(4)	6(2)	11(7)	23(15)	21(9)	33(33)	20(10)	21(8)
	EUR/USD	I	6(5)	5(5)	7(5)	–	10(5)	9(5)	11(5)	8(5)	3(2)	2(2)	4(2)	1(1)
		II	5(0)	6(0)	4(0)	11(11)	1(0)	2(0)	–	3(0)	8(7)	9(7)	7(7)	10(10)
	EUR/GBP	I	10(1)	9(1)	4(1)	1(1)	6(1)	5(1)	7(6)	3(1)	11(1)	–	2(1)	8(1)
		II	1(0)	2(0)	7(1)	10(1)	5(0)	6(1)	4(0)	8(0)	–	11(11)	9(1)	3(1)
MAFE	GBP/USD	I	6(3)	5(3)	9(5)	1(0)	8(5)	10(5)	7(2)	4(2)	2(2)	–	3(2)	11(6)
		II	5(0)	6(1)	2(0)	10(10)	3(0)	1(0)	4(0)	7(4)	9(4)	11(10)	8(4)	0(0)
	Total	I	22(9)	19(9)	20(11)	2(1)	24(11)	24(11)	25(13)	15(8)	16(5)	2(2)	9(5)	20(8)
		II	11(0)	14(1)	13(1)	31(22)	9(0)	9(1)	8(0)	18(4)	17(11)	31(28)	24(12)	13(11)
	EUR/USD	I	11(5)	9(5)	6(5)	1(0)	10(9)	7(5)	8(5)	5(5)	4(2)	3(2)	2(2)	–
		II	–	2(0)	5(1)	10(10)	1(0)	4(1)	3(1)	6(1)	7(7)	8(7)	9(7)	11(10)
	EUR/GBP	I	10(5)	11(6)	7(4)	2(1)	9(7)	6(3)	8(4)	–	3(1)	1(0)	4(2)	5(3)
		II	1(0)	–	4(2)	9(8)	2(0)	5(3)	3(1)	11(0)	8(0)	10(10)	7(7)	6(5)
Qlike	GBP/USD	I	10(4)	6(4)	8(5)	3(2)	9(5)	11(7)	7(5)	–	2(0)	1(0)	4(3)	5(4)
		II	1(0)	5(0)	3(1)	8(8)	2(0)	–	4(1)	11(0)	9(9)	10(9)	7(7)	6(4)
	Total	I	31(14)	26(15)	21(14)	6(3)	28(21)	24(15)	23(14)	5(5)	9(3)	5(2)	10(7)	10(7)
		II	2(0)	7(0)	12(4)	27(26)	5(0)	9(4)	10(3)	28(1)	24(16)	28(26)	23(21)	23(19)

than the HAR models, including CHAR, but still suffer more rejections than the DCS models. The DCS models are successful at forecasting most of the out-of-sample distribution for EUR/USD and GBP/USD. They are less impressive for EUR/GBP, but for no quantile does either UC or UB exceed ten; the same cannot be said for the other models.

7. Conclusion

A score-driven model for RV, formulated with a GB2 distribution in levels, or equivalently an EGB2 distribution in logarithms, is able to capture the main features of a dataset on stock market indices. It is statistically coherent and can approximate long memory, through a two component structure, and asymmetric response (leverage) in a straightforward and transparent manner. The asymptotic theory for the ML estimator in a model with first-order dynamics can be established from the general results in Blasques et al. (2022) and we find that lack of invertibility for models with parameters that tend to arise in practice is unlikely to be a problem. Working in logarithms is useful in that the normal distribution is a limiting case of the balanced EGB2. Furthermore, dynamic heteroscedasticity in the logarithm of RV, sometimes referred to as Volatility of Volatility, can be modeled within the framework of an EGB2 distribution.

A comprehensive forecasting comparison was carried out on two sets of data, one for stocks indices and the other for exchange rates. The best fitting distributions for the GB2-DCS models are the Burr and balanced GB2. The F-distribution also gives a good fit, even though it is not the best. However, it does generalize to the modeling of an $N \times N$ realized volatility covariance matrix as in Opschoor et al. (2018). In all cases the preferred DCS specifications beat all competitors in terms of predictive likelihood, with the inclusion of heteroscedastic effects yielding a slight gain. When other measures are used, no model has a clear forecasting advantage against all the others in terms of point forecasts, but the DCS models produce forecasts that are almost always competitive with the best in any given situation. This remains true even when the HAR model uses additional information, such as series on bi-power variation and semi-variance. The DCS model almost always outperforms the Spline-MEM. The MEM-J fares somewhat better, but it is more computationally intensive.

Accurate modeling of the tails may have important implications for risk management. When the aim is to produce tail forecasts, as reflected in VolaR and ESVol, the flexible fat-tailed GB2 models perform much better than HAR models that are based on an assumption of normality or log-normality. It is also better than the MEM-J, which attempts to capture heavy tails using a mixture of gamma distributions.

Using data on quarticity in modified HAR models offers little or no gain over the DCS and DCS-H models in terms of point forecasts in an exchange rate data set. Furthermore the DCS models forecast VolaR and ESVol much more accurately. This may not be surprising because the starting point of the quarticity models is an asymptotic theory that leads to RV being treated as normally distributed.

Although the proposed GB2 score-driven models contain no major theoretical advances, the results from a comprehensive comparison with a wide range of RV models reinforce the findings in earlier work, such as Catania and Nonejad (2020) and Harvey and Lange (2017, 2018), that, for volatility modeling, score-driven models with two components, short-term leverage effects and a fat-tailed conditional distribution are not only robust, but provide point and density forecasts that are as good as, if not better, than those from other models. There is a strong case for making them the benchmark for others to beat.

Table 8
Results of UC and UB tests for EUR/USD, EUR/GBP and GBP/USD data series.

		p-quantiles	DCS	DCS-H	HAR	HAR in Logs	CHAR	EHAR	HARQ	HARQF	SHARK	RARFIMA	Spline-MEM	MEM-J
		0.1	0.204	1.066	7.558***	191.244***	11.457***	10.846***	7.071***	9.678***	150.282***	0.040	3.455*	5.072**
	UC	0.05	0.096	4.383**	0.291	90.496***	0.427	0.291	0.773	0.773	189.655***	20.246***	0.071	21.166***
		0.01	3.036*	2.303	52.992***	23.955***	45.831***	52.992***	57.966***	63.093***	206.805***	96.809***	13.228***	34.752***
EUR/USD		0.1	0.778	1.489	0.964	−9.643***	0.468	0.774	0.852	0.761	19.663***	−5.586***	−0.261	5.134***
	UB	0.05	0.501	1.368	5.298***	−6.796***	4.734***	5.228***	5.351***	5.407***	22.364***	−1.486	1.823*	7.355***
		0.01	2.250**	2.503**	16.253***	−3.167***	14.585***	15.852***	16.423***	16.928***	29.589***	5.972***	7.647***	6.298***
	Pred Lik		15,230.66	15,234.60	13,931.24	14,682.48	13,933.11	13,927.21	13,899.64	13,880.14	14,081.61	13,430.06	14,983.19	15,008.00
		0.1	3.126*	0.492	55.833***	130.542***	59.003***	54.290***	49.822***	46.976***	58.895***	36.580***	21.122***	11.853***
	UC	0.05	7.493***	5.334**	4.879**	70.085***	4.879**	3.362*	3.362*	2.143	68.316***	125.947***	0.958	15.112***
		0.01	9.173***	6.818***	26.725***	9.976***	24.844***	23.015***	34.752***	39.050***	96.809***	379.475***	10.458***	26.725***
EUR/GBP		0.1	3.159***	2.482**	−2.936***	−8.587***	−3.051***	−2.837***	−2.366**	−2.040**	10.825***	−11.09***	−1.484	5.344***
	UB	0.05	4.140***	3.575***	1.822*	−6.053***	1.736*	1.886*	2.596***	2.977***	12.434***	−7.664***	1.014	6.408***
		0.01	4.307***	3.813***	10.884***	−2.198**	10.814***	10.802***	11.893***	12.675***	17.002***	−3.440***	5.919***	4.899***
	Pred Lik		15,785.69	15,797.29	13,681.72	15,302.16	13,689.38	13,680.21	13,698.66	13,670.19	14,374.99	13,286.91	15,523.04	15,703.12
		0.1	0.013	0.040	42.899***	130.542***	46.976***	42.899***	45.592***	44.233***	55.578***	22.663***	12.086***	1.432
	UC	0.05	1.218	0.773	1.209	67.120***	3.362*	1.489	1.489	1.801	92.773***	69.851***	1.489	8.698***
		0.01	7.959***	5.752**	36.878***	12.488***	34.752***	39.050***	36.878***	41.267***	125.054***	325.860***	16.254***	32.672***
GBP/USD		0.1	0.979	1.048	−1.801*	−8.580***	−2.389**	−1.789*	−1.840*	−1.812*	12.273***	−9.893***	−1.686*	3.235***
	UB	0.05	2.301**	1.686*	2.943***	−6.044***	2.488**	3.046***	3.145***	3.099***	15.009***	−6.475***	1.494	4.933***
		0.01	3.779***	3.606***	15.194***	−2.297**	15.634***	15.496***	15.437***	16.242***	21.444***	−2.331**	7.892***	5.233***
	Pred Lik		15,490.62	15,503.88	12,591.16	15,047.35	12,515.54	12,570.79	12,312.42	12,238.47	12,102.48	12,440.85	15,282.57	15,394.77

* denotes a 10% significance level, ** a 5% significance level and *** a 1% significance level. We also report the predictive likelihood for each model.

Acknowledgments

We would like to thank Siem Jan Koopman, Michael Rockinger, Roberto Casarin and two referees for helpful and constructive comments.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2023.01.029.

References

Andersen, T., Bollerslev, T., Diebold, F., 2007. Roughing it up: Including jump components in the measurement, modeling and forecasting of return volatility. Rev. Econ. Stat. 89 (4), 701–720.
Andersen, T., Bollerslev, T., Diebold, F., Labys, P., 2003. Modeling and forecasting realized volatility. Econometrica 71 (2), 579–625.
Baillie, R., Calonaci, F., Cho, D., Rho, S., 2019. Long memory, realized volatility and heterogeneous autoregressive models. J. Time Series Anal. (4), 609–628.
Barndorff-Nielsen, O., Shephard, N., 2004. Power and bipower variation with stochastic volatility and jumps. J. Financ. Econom. 2 (1), 1–37.
Blasques, F., van Brummelen, J., Koopman, S., Lucas, A., 2022. Maximum likelihood estimation for score-driven models. J. Econometrics 227 (2), 325–346.
Bollerslev, T., Patton, A., Quaedvlieg, R., 2016. Exploiting the errors: A simple approach for improved volatility forecasting. J. Econometrics 192 (1), 1–18.
Brownlees, C., Gallo, G., 2009. Comparison of volatility measures: a risk management perspective. J. Financ. Econom. 8 (1), 29–56.
Buccheri, G., Corsi, F., 2019. HARK the SHARK: Realized volatility modeling with measurement errors and nonlinear dependencies. J. Financ. Econom. 19 (4), 614–649.
Caivano, M., Harvey, A., 2014. Time series models with an EGB2 conditional distribution. J. Time Series Anal. 34, 558–571.
Caporin, M., Rossi, E., Santucci de Magistris, P., 2017. Chasing volatility: A persistent multiplicative error model with jumps. J. Econometrics 198 (1), 122–145.
Catania, L., Nonejad, N., 2020. Density forecasts and the leverage effect: Evidence from observation and parameter-driven volatility models. Eur. J. Finance 26 (2–3), 100–118.
Christoffersen, P., 1998. Evaluating interval forecasts. Internat. Econom. Rev. 39 (4), 841–862.
Corsi, F., 2009. A simple approximate long-memory model of realized volatility. J. Econometrics 7, 174–196.
Corsi, F., Mittnik, S., Pigorsch, C., Pigorsch, U., 2008. The volatility of realized volatility. Econometric Rev. 27 (1), 46–78.
Corsi, F., Renò, R., 2012. Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. J. Bus. Econom. Statist. 30, 368–380.
Creal, D., Koopman, S.J., Lucas, A., 2013. Generalized autoregressive score models with applications. J. Appl. Econometrics 28, 777–795.
Diebold, F., Mariano, R., 2002. Comparing predictive accuracy. J. Bus. Econ. Stat. 20 (1), 134–144.
Du, Z., Escanciano, J., 2017. Backtesting expected shortfall: Accounting for tail risk. Manage. Sci. 63 (4), 940–958.
Harvey, A.C.,
Harvey, A.C., 2022. Score-driven time series models. Annu. Rev. Stat. Appl. 9 (1), 321–342.
Harvey, A., Lange, R.-J., 2017. Volatility modelling with a generalized-t distribution. J. Time Series Anal. 38 (2), 175–190.
Harvey, A., Lange, R.-J., 2018. Modeling the interactions between volatility and returns. J. Time Series Anal. 39 (6), 909–919.
Kleiber, C., Kotz, S., 2003. Statistical Size Distributions in Economics and Actuarial Sciences. Wiley.
Koopman, S., Lit, R., Harvey, A., 2021. STAMP. Structural Time Series Analyser, Modeller and Predictor. Timberlake Consultants Ltd, London.
McDonald, J., Xu, Y., 1995. A generalization of the beta distribution with applications. J. Econometrics 66, 133–152.
Mitchell, J., Wallis, K.F., 2011. Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness. J. Appl. Econometrics 26 (6), 1023–1040.
Müller, U., Dacorogna, M., Dav, R., Olsen, R., Pictet, O., von Weizsacker, J., 1997. Volatilities of different time resolutions - analysing the dynamics of market components. J. Empir. Financ. 4, 213–239.
Opschoor, A., Janus, P., Lucas, A., Van Dijk, D., 2018. New HEAVY models for fat-tailed realized covariances and returns. J. Bus. Econom. Statist. 36 (4), 643–657.
Opschoor, A., Lucas, A., 2023. Time-varying variance and skewness in realized volatility measures. Int. J. Forecast. 39 (2), 827–840.
Patton, A.J., 2011. Volatility forecast comparison using imperfect volatility proxies. J. Econometrics 160 (1), 246–256.
Patton, A., Sheppard, K., 2015. Good volatility, bad volatility: Signed jumps and the persistence of volatility. Rev. Econ. Stat. 97, 683–697.
Taylor, S., 2005. Asset price dynamics, volatility, and prediction.