Quantitative Finance

ISSN: 1469-7688 (Print) 1469-7696 (Online) Journal homepage: http://www.tandfonline.com/loi/rquf20

Semi-parametric Bayesian tail risk forecasting incorporating realized measures of volatility

Richard Gerlach, Declan Walpole & Chao Wang

To cite this article: Richard Gerlach, Declan Walpole & Chao Wang (2017) Semi-parametric Bayesian tail risk forecasting incorporating realized measures of volatility, Quantitative Finance, 17:2, 199-215, DOI: 10.1080/14697688.2016.1192295

To link to this article: http://dx.doi.org/10.1080/14697688.2016.1192295

Published online: 15 Jul 2016.

Quantitative Finance, 2017
Vol. 17, No. 2, 199–215, http://dx.doi.org/10.1080/14697688.2016.1192295

Semi-parametric Bayesian tail risk forecasting incorporating realized measures of volatility

RICHARD GERLACH*† , DECLAN WALPOLE‡ and CHAO WANG†

†Discipline of Business Analytics, The University of Sydney, Sydney, Australia
‡Deposits and Transactions Analytics, Commonwealth Bank of Australia, Sydney, Australia

(*Received 1 November 2015; accepted 12 May 2016; published online 15 July 2016*)

Realized measures employing intra-day sources of data have proven effective for dynamic volatility and tail-risk estimation and forecasting. Expected shortfall (ES) is a tail risk measure, now recommended by the Basel Committee, involving a conditional expectation that can be semi-parametrically estimated via an asymmetric sum of squares function. The conditional autoregressive expectile class of model, used to implicitly model ES, has been extended to allow the intra-day range, not just the daily return, as an input. This model class is here further extended to incorporate information on realized measures of volatility, including realized variance and realized range (RR), as well as scaled and smoothed versions of these. An asymmetric Gaussian density error formulation allows a likelihood that leads to direct estimation and one-step-ahead forecasts of quantiles and expectiles, and subsequently of ES. A Bayesian adaptive Markov chain Monte Carlo method is developed and employed for estimation and forecasting. In an empirical study forecasting daily tail risk measures in six financial market return series, over a seven-year period, models employing the RR generate the most accurate tail risk forecasts, compared to models employing other realized measures as well as to a range of well-known competitors.

Keywords: CARE model; Realized measures; Expected shortfall; Expectiles; Markov chain Monte Carlo method

JEL Classification: C11, C22, C53, C58

1. Introduction

Daily Value-at-Risk (VaR) and expected shortfall (ES; Artzner *et al.* 1999), also known as conditional VaR (CVaR), are two of the main tools employed for tail risk measurement and capital allocation by financial institutions, in order to comply with their requirements under the current Basel Capital Accord. ES is now favoured over VaR by the Accord, in its latest incarnation, mainly since VaR is not a mathematically coherent measure i.e. it is not ‘sub-additive’ (see Artzner *et al.* 1997, 1999) and since VaR does not give a measure of the expected loss in the event of an extreme return; ES overcomes both these negative aspects. ES is the expected return in the part of the return distribution that is more extreme than a given quantile and hence is also called conditional VaR. ES was proposed in May 2012 by the Basel Committee on Banking Supervision, in their ‘fundamental review of the trading book—consultative document’ (http://www.bis.org/publ/bcbs219.htm) in preference to VaR, since it is ‘a risk measure that better captures tail risk’. The purpose of this paper is to extend a class of semi-parametric tail risk models, so as to incorporate realized measures of volatility, to more accurately forecast both ES and VaR in daily financial asset return data.

Taylor (2008) proposed a semi-parametric method of estimation for VaR and ES based on the theory of expectiles, a quantity defined by Aigner *et al.* (1976) as the minimum of the expectation of an asymmetric sum of squares, and also called a partial moment by those authors. Taylor (2008) subsequently proposed the conditional auto-regressive expectile (CARE) class of model, estimated by Asymmetric Least Squares (ALS), then employed a connection between quantiles, expectiles and ES, based on a result found by Newey and Powell (1987), to subsequently estimate and forecast both VaR and ES simultaneously. Standard CARE models employ squared or absolute daily returns as an explanatory input variable. Gerlach and Chen (2016) extend the class of CARE models to incorporate information from the intra-day range, see Parkinson (1980), Garman and Klass (1980), and further, develop a Bayesian estimator. Gerlach and Chen (2016) find the Bayesian estimator to be more efficient in estimation and more accurate in VaR and ES forecasting, compared to ALS. As such, their estimator is employed in this paper.

*Corresponding author. Email: richard.gerlach@sydney.edu.au

Several realized volatility measures are known to be more efficient than daily close-close returns, which can miss large intra-day movements, and intra-day range: e.g. realized variance, Andersen and Bollerslev (1998) and Alizadeh *et al.* (2002); and realized range (RR), Martens and van Dijk (2007) and Christensen and Podolskij (2007). The latter measure employs the sum of high-frequency intra-period squared ranges, simulations in Martens and van Dijk (2007) showed that it compared favourably to realized volatility (RV) in volatility estimation across various micro-structure noise settings. This paper further extends the CARE class to include a general realized measure, and investigates the forecasting performance of several recent, modern and competing measures, including RV and RR, when incorporated into the CARE framework. Scaling, as in Martens and van Dijk (2007), and smoothing by sub-sampling, as in Zhang *et al.* (2005), are also considered for both RV and RR to produce possibly smoother and less biased measures.

This paper extends the CARE model class to incorporate any observed, explanatory input variable. The proposed CARE-X models, and the MCMC estimation methods employed, are examined through application to various financial market stock index returns in a study of one-step-ahead ES forecasting. This illustrates that CARE-X models out-perform both existing CARE models and a range of competing models and methods, over the forecast period 2008–2014. The tail risk forecast combination methods of Chang *et al.* (2011) and McAleer *et al.* (2013) are also incorporated into this study.

The rest of the paper is organized as follows: section 2 reviews expectiles, their estimation and their general link with quantiles and ES, then reviews the CARE models in Taylor (2008) and Gerlach and Chen (2016); section 3 discusses realized measures and introduces the proposed CARE-X models; section 4 presents a likelihood formulation and the adaptive MCMC methods employed for estimation; section 5 discusses assessing ES forecasts and section 6 contains the empirical study. Concluding remarks are in section 7.

2. Expectiles, care models and ES

This section reviews the definition of an expectile, existing CARE models and discusses the link between expectiles and ES.

2.1. Expectiles and ES

Aigner *et al.* (1976) defined the τ-level expectile (or 'partial moment') for a continuous random variable (r.v.) Y as the value of μ_τ that minimizes the expectation $E[|\tau - I[Y < \mu_\tau]|(Y - \mu_\tau)^2]$; where $\tau \in (0, 1)$ and $I(.)$ is an indicator function that equals one when Y is less than the expectile μ_τ, and equals zero otherwise. Note that $\mu_{0.5} = E(Y)$.

Based on a sample $y_1, \ldots, y_T$ on Y, and a fixed, known τ, the constant τ-level expectile of Y, denoted μ_τ, can be estimated by minimizing the asymmetric sum of squares function:

$$\sum_{t=1}^{T} |\tau - I[y_t < \mu_\tau]|(y_t - \mu_\tau)^2. \tag{1}$$

This is called the ALS estimator.

ES is defined as the expected value of an r.v. Y, conditional on Y being more extreme than its α-level quantile: i.e. $\text{ES}_\alpha = E(Y|Y < Q_\alpha)$, where Q_α is the quantile of Y. Here we consider only $\alpha < 0.5$ and thus restrict this work to left-tail or negative risk on long positions, as standard in this literature.

Newey and Powell (1987) found a one-to-one connection between expectiles, quantiles and conditional expectations in general. As presented in Taylor (2008), if $E(Y) = 0$ then this relationship can be written, in terms of ES, as:

$$\text{ES}_\alpha = \left(1 + \frac{\tau}{(1-2\tau)\alpha}\right)\mu_\tau\,, \tag{2}$$

where most terms are defined above. Here τ is specifically chosen so that the τ-level expectile μ_τ coincides with the α-level quantile Q_α of Y. This is always possible theoretically, since the expectile μ_τ must occur at some quantile of the distribution of Y. Taylor (2008) showed that both (1) and (2) also apply in the case of parameter estimation in a time-varying conditional expectile $\mu_{\tau;t}$ model, e.g. the CARE model discussed in the next section; whereas the results in Newey and Powell (1987) are only for constant or unconditional expectiles. τ is subsequently removed from the subscript for brevity.

2.2. CARE modelling

Engle and Manganelli (2004) discussed the three CAViaR models: symmetric absolute value (SAV), asymmetric (AS) and indirect GARCH (IG). Taylor (2008) took the same forms as these models, replacing the dynamic quantile (DQ) with dynamic expectile terms and proposed three models: CARE-SAV, CARE-AS and CARE-IG, with the following specifications (where y_t is a financial return at time t):

CARE-SAV:

$$\mu_t = \beta_1 + \beta_2\mu_{t-1} + \beta_3\,|y_{t-1}|\,, \tag{3}$$

where the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ are unrestricted on the real line;

CARE-AS:

$$\mu_t = \beta_1 + \beta_2\mu_{t-1} + (\beta_3 I_{[y_{t-1}>0]} + \beta_4 I_{[y_{t-1}<0]})|y_{t-1}|\,, \tag{4}$$

where the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ are unrestricted on the real line and where the conditional expectile responds asymmetrically to positive and negative returns; and

CARE-IG:

$$\mu_t = -\sqrt{\beta_1 + \beta_2\mu_{t-1}^2 + \beta_3 y_{t-1}^2}, \tag{5}$$

where all parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ are restricted to be greater than 0, ensuring positivity under the square root.

The intra-day range is employed in a CARE-R model developed by Gerlach and Chen (2016). The standard definition of the intra-day range is:

$$R_t = (\log(H_t) - \log(L_t)) \times 100$$

where H_t is the highest price or index value, and L_t is the lowest price or index value, during day t. Gerlach and Chen (2016) incorporated overnight price movements into this measure, defining range plus overnight as:

$$RaO_t = (\log(\max(C_{t-1}, H_t)) - \log(\min(C_{t-1}, L_t)))\times 100\ ,$$

where C_t is the closing price on day t.

Two CARE-R models developed by Gerlach and Chen (2016) are:

Range CARE-SAV (CARE-R-SAV):

$$\mu_t = \beta_1 + \beta_2\mu_{t-1} + \beta_3 R_{t-1}, \tag{6}$$

where the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ are unrestricted on the real line.

Range Indirect-GARCH (CARE-R-IG):

$$\mu_t = -\sqrt{\beta_1 + \beta_2\mu_{t-1}^2 + \beta_3 R_{t-1}^2}. \tag{7}$$

Here all parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ are restricted greater than 0 to ensure positivity under the square root.

These models can all produce one-step-ahead (only) forecasts of μ_t (expectiles). Whilst stationarity conditions have not been theoretically considered in the literature, it is logical that a necessary condition would be $\beta_2 < 1$, so that μ_t did not diverge; but this is not a sufficient condition for stationarity.

3. Realized measures and CARE-X models

This section gives a brief introduction to various realized measures of volatility considered here and presents the proposed CARE-X models.

3.1. *Realized measures*

First, for day t, denote the intra-day high, low and closing prices as H_t, L_t and C_t. The daily percentage log return is then:

$$r_t = 100[\log(C_t) - \log(C_{t-1})] \tag{8}$$

Assuming the mean return is zero, as standard, a constant daily return variance can be estimated from a sample $r_1, \ldots, r_T$ by:

$$V_t = \frac{1}{T}\sum_{i=0}^{T-1} r_{t-i+1}^2 \tag{9}$$

The typical squared daily return measure takes $T = 1$.

Based on the distribution of range derived by Feller (1951), Parkinson (1980) proposed the high-low intra-day range (squared), with scaling factor $4\log 2$ as an approximately unbiased variance estimator:

$$Ra_t^2 = \frac{(\log H_t - \log L_t)^2}{4\log 2} \tag{10}$$

The equivalent range plus overnight variance estimator, RaO_t is calculated by substituting $\max(C_{t-1}, H_t)$ for H_t and $\min(C_{t-1}, L_t)$ for L_t. Through theoretical derivation and a simulation study, Parkinson showed that Ra_t is a more efficient estimator than the traditional squared return. Garman and Klass (1980), Rogers and Satchell (1991) and Yang and Zhang (2000) derived other range-based estimators; a full study and comparison on the properties of different volatility estimators is presented in Molnár (2012).

Extending into the high-frequency intra-day framework, each day t can be divided into N equally sized intervals of length $\triangle$, each intra-day time subscripted as $i = 0, 1, 2, \ldots, N$. The log closing price at the ith interval of day t is denoted as $P_{t-1+i\triangle}$. Then, the high and low prices during this time interval are $H_{t,i} = \sup_{(i-1)\triangle<j<i\triangle} P_{t-1+j}$ and $L_{t,i} = \inf_{(i-1)\triangle<j<i\triangle} P_{t-1+j}$, respectively. Realized variance (RV) has proven an efficient volatility estimator and gained popularity in recent years. RV is simply the sum of the N intra-day squared returns, at frequency $\triangle$, for day t, i.e.

$$\mathrm{RV}_t^{\triangle} = \sum_{i=1}^{N}[log(P_{t-1+i\triangle}) - log(P_{t-1+(i-1)\triangle})]^2 \tag{11}$$

The RR, proposed by Martens and van Dijk (2007) and Christensen and Podolskij (2007), has the following specification, which simply replaces the intra-day squared returns with intra-day squared ranges, and scales as in Parkinson (1980):

$$\mathrm{RR}_t^{\triangle} = \frac{\sum_{i=1}^{N}(\log H_{t,i} - \log L_{t,i})^2}{4\log 2} \tag{12}$$

Theoretically, the RR may contain more information about volatility than RV, in the same way that the intra-day range contains more information than squared returns, i.e. it uses all the price movements in a time period, not just the price at the start and end of each sub-period. Results in Martens and van Dijk (2007) lend support to this hypothesis.

Of course, both RV and RR have been criticized as being subject to micro-structure noise causing bias and inefficiency, more so than daily returns or daily ranges. This issue has been studied extensively, see Rogers and Satchell (1991), Barndorff-Nielsen *et al.* (2004) and Christensen and Podolskij (2007) for discussion. In response, Martens and van Dijk (2007) presented a scaling process, as in equations (13) and (14).

$$RVSc_t^{\triangle} = \frac{\sum_{l=1}^{q}\mathrm{RV}_{t-1}}{\sum_{l=1}^{q}\mathrm{RV}_{t-1}^{\triangle}}\mathrm{RV}_t^{\triangle}, \tag{13}$$

$$\mathrm{RRSc}_t^{\triangle} = \frac{\sum_{l=1}^{q}\mathrm{RR}_{t-1}}{\sum_{l=1}^{q}\mathrm{RR}_{t-1}^{\triangle}}\mathrm{RR}_t^{\triangle}, \tag{14}$$

where RV_{t-1} and RR_{t-1} represent the daily return square and range square at day $t-1$ and q is the number of days employed to estimate the scaling factors. This scaling process is motivated by the fact that the daily return and range are less affected by micro-structure noise and thus can be used to help reduce bias.

Further, Zhang *et al.* (2005) proposed a sub-sampling process to further smooth out micro-structure noise. For day t, N equally sized samples are grouped into M non-overlapping subsets $X^{(m)}$ with size $N/M = n_k$, which means:

$$X = \bigcup_{m=1}^{M} X^{(m)}, \text{ where } X^{(k)} \cap X^{(l)} = \emptyset, \text{ when } k \neq l. \tag{15}$$

Then sub-sampling will be implemented on the subsets $X^{(i)}$ with n_k interval:

$$X^{(i)} = i, i+n_k, \ldots, i+n_k(M-2), i+n_k(M-1), \\ \text{where } i = 0, 1, 2\ldots, n_k - 1. \tag{16}$$

Representing the log closing price at the ith interval of day t as $C_{t,i} = P_{t-1+i\triangle}$, the RV with the subsets X^i is:

$$\mathrm{RV}_i = \sum_{m=1}^{M}(C_{t,i+n_k m} - C_{t,i+n_k(m-1)})^2; \\ \text{where } i = 0, 1, 2\ldots, n_k - 1. \tag{17}$$

We have the T/M RV with T/N sub-sampling as (supposing there are T min per trading day):

$$\mathrm{RVSS}_{T/M,T/N} = \frac{\sum_{i=0}^{n_k-1} \mathrm{RV}_i}{n_k}, \tag{18}$$

Then, denoting the high and low prices during the interval $i + n_k(m-1)$ and $i + n_k m$ as $H_{t,i} = \sup_{(i+n_k(m-1))\Delta<j<(i+n_k m)\Delta} P_{t-1+j}$ and $L_{t,i} = \inf_{(i+n_k(m-1))\Delta<j<(i+n_k m)\Delta} P_{t-1+j}$, respectively, we propose the T/M RR with T/N sub-sampling as:

$$\mathrm{RR}_i = \sum_{m=1}^{M} (H_{t,i} - L_{t,i})^2; \quad \text{where } i = 0, 1, 2\ldots, n_k - 1. \tag{19}$$

$$\mathrm{RRSS}_{T/M,T/N} = \frac{\sum_{i=0}^{n_k-1} \mathrm{RR}_i}{4\log 2 n_k}, \tag{20}$$

For example, the 5 min RV and RR with 1 min sub-sampling can be calculated, respectively, as below :

$$\begin{aligned}
\mathrm{RV}_{5,1,0} &= (\log C_{t5} - \log C_{t0})^2 \\
&\quad + (\log C_{t10} - \log C_{t5})^2 + \cdots \\
\mathrm{RV}_{5,1,1} &= (\log C_{t6} - \log C_{t1})^2 \\
&\quad + (\log C_{t11} - \log C_{t6})^2 + \cdots \\
\mathrm{RVSS}_{5,1} &= \frac{\sum_{i=0}^{4} \mathrm{RV}_{5,1,i}}{5} \\
\mathrm{RR}_{5,1,0} &= (\log H_{t0\le t\le t5} - \log L_{t0\le t\le t5})^2 \\
&\quad + (\log H_{t5\le t\le t10} - \log L_{t5\le t\le t10})^2 + \cdots \\
\mathrm{RR}_{5,1,1} &= (\log H_{t1\le t\le t6} - \log L_{t1\le t\le t6})^2 \\
&\quad + (\log H_{t6\le t\le t11} - \log L_{t6\le t\le t11})^2 + \cdots \\
\mathrm{RRSS}_{5,1} &= \frac{\sum_{i=0}^{4} \mathrm{RR}_{5,1,i}}{4\log(2)5}
\end{aligned}$$

Only intra-day returns on the 5 min frequency, additionally with 1 min sub-sampling when employed, are considered in this paper.

3.2. CARE-X models

Each realized measure discussed, and others not considered here, may be input as an explanatory variable into two general CARE-X specifications now presented. Denote the realized measure on day t as X_t. Then, the two specifications are:

Realized CARE-SAV (CARE-X-SAV):

$$\mu_t = \beta_1 + \beta_2 \mu_{t-1} + \beta_3 X_{t-1}, \tag{21}$$

where the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ are unrestricted on the real line.

Realized Indirect-GARCH (CARE-X-IG):

$$\mu_t = -\left[\beta_1 + \beta_2 \mu_{t-1}^2 + \beta_3 X_{t-1}^2\right]^{1/2}. \tag{22}$$

These models again produce one-step-ahead (only) forecasts of μ_t (expectile). It is again logical that a necessary condition for stationarity would be $\beta_2 < 1$, so that μ_t does not diverge. For the IG model, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ are further restricted greater than 0, to ensure positivity under the square root. No other parameter restrictions are enforced.

Other CARE-X models are possible, e.g. CARE-X-AS, where asymmetry in expectile could be driven by returns (sign asymmetry) or by the realized measure (size asymmetry; see Chen *et al.* 2008), however these are left for future work.

4. ALS, Likelihood and Bayesian estimation

Taylor (2008) estimated the CARE model unknown parameters $\boldsymbol{\beta}$ by ALS, as in (1). The method relies on a grid search to estimate τ, where at each value of τ, $\boldsymbol{\beta}$ is estimated by ALS conditional on that value and the optimal value of τ is chosen to make the violation rate of μ_t as close as possible to α; we follow the same approach to estimate τ and condition the Bayesian estimator on that estimate.

Since the expectile series in a CARE model are chosen to coincide with the α-level quantile series, quantile regression or CaViaR (Engle and Manganelli 2004) methods of estimation could be used to estimate CARE models. However, Taylor (2008) showed the increase in both efficiency of estimation and in speed of optimization of ALS compared to these methods.

Bayesian methods generally require the specification of a likelihood function and a prior distribution. Gerlach and Chen (2016) considered an asymmetric Gaussian (AG) density, allowing a likelihood formulation and Bayesian estimation for CARE models. Consider a sample of daily returns, $y_1, \ldots, y_T$ and associated realized measures.

Gerlach and Chen (2016) write the general model as:

$$y_t = \mu_t + \epsilon_t, \quad \epsilon_t \sim \mathrm{AG}(\tau, 0, \sigma), \tag{23}$$

where ϵ_t is an i.i.d. AG process with mode 0, shape parameter τ and scale σ. The kernel of a pdf for an AG r.v. is defined be the exponential of the negative of the ALS function (1), with an added scale factor σ. The nuisance parameter σ can subsequently be analytically integrated out, here under a standard, conjugate Jeffreys prior, since the likelihood constructed is proportional to an inverse gamma density in σ. The corresponding integrated, or marginal, likelihood function, conditional on τ, is:

$$L_\tau(\mathbf{y}|\boldsymbol{\beta}, \tau) \propto \left(\sum_{t=1}^{T} (y_t - \mu_t)^2 \left|\tau - I[y_t < \mu_t(\boldsymbol{\beta})]\right| \right)^{-T/2} . \tag{24}$$

Because of the manner in which (1) is contained inside (24), the maximum-likelihood (ML) estimation for $\boldsymbol{\beta}$, is equivalent to the ALS estimator from (1).

Even if the assumption of AG errors were accurate and reasonable, the methods discussed here can be used only to produce one-step-ahead forecasts of expectiles and ES. This is because the realized measures are treated as exogenous variables that must be known as an input in the CARE-X model class. To produce a two-step ahead forecast of ES, a model for the one-step-ahead realized measure would have to be appended to the CARE-X model, so that future realizations of the realized measure could be generated. We do not consider such models here and leave this for future work.

The likelihood for all the CARE and CARE-X family of models is completely specified by (24), combined with the relevant formula for $\mu_t(\boldsymbol{\beta})$ from (3)–(22).

Gerlach and Chen (2016) found their Bayesian MCMC estimator to be more efficient in estimation and more accurate in VaR and ES forecasting, compared to ALS, for a CARE-X model employing the range as an input. Their MCMC sampler is adapted for general CARE-X models. First the prior distribution required for a Bayesian approach is specified.

4.1. Prior and posterior densities

The prior is chosen to be close to uninformative over the possible region for the parameters $\boldsymbol{\beta}$, with two exceptions: (i) A Jeffreys prior for the scale parameter σ in the AG distribution, and (ii) a shrinkage prior for the intercept parameter in each model, as in Gerlach and Chen (2016) i.e.

$$\pi(\boldsymbol{\beta}, \sigma) \propto I(A)\frac{1}{\sigma}\frac{1}{\beta_1},$$

for the CARE-X-IG and CARE-X-SAV models. This is a mostly flat prior on the parameters in $\boldsymbol{\beta}$, restricted by the indicator function being non-zero only over the region A. A includes the restrictions necessary for stationarity, and sufficient to ensure positivity under the square root for IG-type models, as discussed above. Whilst the infinite regions A mean the prior is improper for each model, the likelihood, in a form similar to a Student-t density with $\sim T$ degrees of freedom in $\boldsymbol{\beta}$, restricted to A, is thus clearly integrable in terms of $\boldsymbol{\beta}$, and hence the resulting posterior is proper.

Using equation (24) and any of (3)–(22), plus the priors above, the joint posterior density for $\boldsymbol{\beta}, \sigma | \mathbf{y}, \tau$ is simply proportional to the likelihood times the prior, evaluated over the region defined by A. This posterior is not in the form of a known distribution in $\boldsymbol{\beta}$, computational methods are thus employed for estimation.

4.2. Adaptive MCMC sampling using Metropolis methods

Sampling from $p(\boldsymbol{\beta}|\mathbf{y}, \tau)$ directly is not possible, so a dependent Markov chain Monte Carlo (MCMC) sample is obtained from $\boldsymbol{\beta}|\mathbf{y}, \tau$ via adaptive versions of the Metropolis and Metropolis–Hastings (MH) (Metropolis *et al.* 1953, Hastings 1970) algorithms. A similar sampling algorithm to that in Gerlach and Chen (2016) is employed, utilizing a random walk Metropolis (RW-M) algorithm during the burn-in period and an independent kernel (IK-) MH algorithm during the sampling period. The original idea comes from Chen and So (2006), though we adapt it. The diagonal scale matrix used in the burn-in period is tuned to achieve optimal acceptance rates, i.e. 40% for single parameters and 24% for groups of parameters of dimension > 1, based on recommendations in Gelman *et al.* (1996) and Chen and So (2006).

We employ Student-t proposals, with low degrees of freedom, in the burn-in period. That is, a RW-M method, using a multivariate Student-$t(5)$ distribution, with a diagonal covariance matrix, is used as the proposal distribution in the burn-in. This method is a special simplified case of the more general and flexible 'AdMit' mixture of Student-t proposal procedure proposed by Hoogerheide *et al.* (2007). The MCMC sampling period employs an IKMH algorithm, employing the sample mean ($\bar{X}$) and sample covariance matrix (S) of the burn-in period iterates as the mean and scale matrix for the proposal distribution in the IKMH algorithm. The chosen proposal is a mixture of three Gaussian distributions, each with mean $\bar{X}$ and var-cov matrix $C_i S$, $i = 1, \ldots, 3$, where $C_1 = I$, $C_2 = 10I$ and $C_3 = 100I$. This mixture induces fat-tails into the proposal distribution and helps to ensure the algorithm does not get stuck. Posterior means, using only iterates from the sampling period, are employed to provide estimates of parameters.

5. Assessing tail risk forecasts

For CARE-type models, once the MCMC samples for $\boldsymbol{\beta}$ are obtained for a particular CARE model and data-set, one-step-ahead forecasts of the relevant expectile (also the quantile) can be obtained via the dynamic expectile CARE equation, and then equation (2) used to form a one-step ahead forecast of conditional ES. Under a Bayesian framework, each parameter vector iterate in the sampling period is employed together with the CARE equation and equation (2) to form an MCMC iterate of a one-step ahead forecast of conditional ES and of VaR. These forecasts are averaged to form posterior mean one-step ahead forecasts of ES and VaR for each day in the forecast period.

Various common tests can be applied to directly assess VaR quantile forecasts: e.g. the unconditional coverage (UC), independence of violations (IND) and conditional coverage (CC) tests of Kupiec (1995) and Christoffersen (1998), respectively, as well as the DQ test of Engle and Manganelli (2004) and VQR test of Gaglianone *et al.* (2011); these are all applied here. These tests are now standard and we refer readers to the original papers for details.

Proper or optimal assessment of a set of ES forecasts is still an issue under investigation in the literature. We consider and discuss several approaches for this.

5.1. Loss functions

Cost or loss measures can be applied to assess ES forecasts, as in So and Wong (2012) who employed RMSE and MAD of the 'ES residuals' $y_t - \mathrm{ES}_t$, only for days when the return violates the associated VaR forecast, i.e. $y_t < \mathrm{VaR}_t$. However, these loss functions are not minimized by the true ES series; Gneiting (2011) showed that ES is not 'elicitable': i.e. there is no loss function that is minimized by the true ES series, in general. Recently, however Fissler and Ziegel (in press) developed a family of loss functions, that are a joint function of the associated VaR and ES series, that are minimized by the true VaR and ES series, i.e. they are strictly consistent scoring functions for (VaR, ES). The function is of the form:

$$\begin{aligned} S_t(y_t, \mathrm{VaR}_t, \mathrm{ES}_t) &= (I_t - \alpha)G_1(\mathrm{VaR}_t) - I_t G_1(y_t) + G_2(\mathrm{ES}_t) \\ &\quad \times \left(\mathrm{ES}_t - \mathrm{VaR}_t + \frac{I_t}{\alpha}(\mathrm{VaR}_t - y_t)\right) \\ &\quad - H(\mathrm{ES}_t) + a(y_t), \end{aligned}$$

where $I_t = 1$ if $y_t < \mathrm{VaR}_t$ and 0 otherwise for $t = 1, \ldots, T$, $G_1()$ is increasing, $G_2()$ is strictly increasing and strictly convex, $G_2 = H^{'}$ and $\lim_{x \to -\infty} G_2(x) = 0$ and $a(\cdot)$ is a real-valued integrable function. Motivated by a suggestion in Fissler *et al.* (2015), making the choices: $G_1(x) = x$, $G_2(x) = exp(x)$, $H(x) = exp(x)$ and $a(y_t) = 1 - \log(1 - \alpha)$, which satisfy the required criteria, returns the scoring function:

$$S_t(y_t, \mathrm{VaR}_t, \mathrm{ES}_t) = (I_t - \alpha)\mathrm{VaR}_t - I_t y_t + \exp(\mathrm{ES}_t) \times \left(\mathrm{ES}_t - \mathrm{VaR}_t + \frac{I_t}{\alpha}(\mathrm{VaR}_t - y_t)\right) - \exp(\mathrm{ES}_t) + 1 - \log(1-\alpha), \quad (25)$$

where the loss function is $S = \sum_{t-1}^{T} S_t$. Here, S is a strictly consistent scoring rule that is jointly minimized by the true VaR and ES series; we use this to informally and jointly assess and compare the VaR and ES forecasts from all models.

5.2. *Testing ES residuals*

The most common formal testing method for ES forecast accuracy is based on the fact that ES is a conditional expectation beyond a VaR quantile. An informal assessment examines the ES residuals for data that violate the corresponding quantile VaR predictions, i.e. where $y_t < \mathrm{VaR}_t$ and $I_t = 1$. This includes whether these ES residuals have mean close to 0, or not; as they should if the VaR and ES forecasts are accurate.

Since financial returns are usually not i.i.d, the ES residuals are often scaled by a measure of predicted volatility, e.g. see McNeil and Frey (2000), or by the predicted VaR levels, as in Taylor (2008). The latter approach is taken here, since CARE models do not provide a volatility estimate. A formal approach then tests for a non-zero mean of these standardized ES residuals, $\frac{y_t - \mathrm{ES}_t}{\mathrm{VaR}_t}$ during the forecast period. Subsequently, the standard non-parametric bootstrap test of Efron and Tibshirani (1993), as employed in Taylor (2008) and So and Wong (2012), is applied to the standardized ES residual series, $\frac{y_t - \mathrm{ES}_t}{\mathrm{VaR}_t}$, for days in the forecast period where the return violates the VaR threshold. This test is applied to assess the competing ES forecasting models. However, this test may not have optimal power, due to the low sample sizes of returns that are more extreme than the VaR predictions, i.e. $\approx n\alpha$ residuals under the null hypothesis. We have confirmed this in a simulation study, whose results are available from the authors on request and will appear in future work.

5.3. *Alternative ES testing methods*

Several other ES back-testing procedures have been proposed, e.g. tests by Wong (2008), Costanzino and Curran (2015), Du and Escanciano (2015) and Acerbi and Szekeley (2014); however, all these tests require the forecast model to produce an estimate of the forecast cumulative distribution function (cdf), at least in the tail of the distribution, i.e. beyond the VaR estimate. Unfortunately, CARE-type models do not produce this cdf estimate, and thus these tests cannot be employed in this paper.

As an alternative, Kerkhof and Melenberg (2004) suggested comparing ES, and other risk measurement methods, on an equal quantile basis, in the same way VaR models are back-tested by their violation rate. This method relies on calculating or approximating the specific quantile level that the ES falls at and has been applied to fully parametric ES models (e.g. GARCH), where the exact ES quantile level can be calculated, see e.g. Chen *et al*. (2012). In that case, the UC, CC, DQ and VQR tests can be applied using the quantile level of the ES, treating the ES series as a VaR prediction. Naturally this approach assumes that the ES falls at a consistent quantile level of the conditional distribution of the series being analysed, which may seem like a strong assumption; e.g. it assumes that if Student-*t* errors were fit then the degrees of freedom were constant over the forecast period. However, table 1 shows the ES values and their quantile levels for various distributions and indicates a surprising result.

First, the ES values differ substantially across distributional choices and within distributions for differing parameter values. However, the quantile levels that ES occurs at differ far less over the same settings. Chen and Gerlach (2013) adapted the two-sided Weibull (TW) distribution of Malevergne and Sornette (2004) as an error distribution in a GARCH model. When applied to ten financial time series, the estimated conditional TW distribution had δ_α values consistently in $(0.0035 - 0.0037)$ for $\alpha = 0.01$. These ES quantile levels, denoted δ_α in table 1, do vary across different distributions or different degrees of freedom for a Student-*t* or skewed-*t*, however, the variation is quite small. As pointed out by Gerlach and Chen (2016): (i) distributions with fatter tails tend to have ES fall at slightly lower quantiles, as expected; (ii) the Gaussian's tails are well known as too thin for return data, thus ES should fall at lower quantile levels than the Gaussian suggests; (iii) for most real (daily) return data-sets, the degrees of freedom for a Student-*t* or skewed-*t* is estimated between 6 and 15: for the distributions here, we expect ES to fall between the quantile levels (0.0034, 0.0037) for a 1% ES for these distributions; (iv) finally, when examining forecast periods of returns of between 100 and 1000 days, as is most common, we see that any test applied should be robust, or insensitive, to a choice of quantile level within these ranges, i.e. the ranges in table 1 are quite small and narrow for practical forecast data sizes and the tests applied to them; i.e. most tests of forecast accuracy will not be able to distinguish between violation rates in these ranges at the typical sample sizes used. This includes the case where the shape of the distribution changes over time, e.g. a Student-*t* changing degrees of freedom anywhere from 6 to ∞, the shape changing from Student-*t* to AL or TW and/or back again; table 1 illustrates that these sorts of changes will barely move the quantile level that ES falls at. See Gerlach and Chen (2016) for an extended discussion.

As such, we agree with Gerlach and Chen (2016) that it is sensible for CARE-type, and other semi- or non-parametric models, to also assume that the 1% ES values will occur approximately at the 0.36% quantile level, and to use that quantile level to assess the CARE model (and non-parametric) forecasts of ES using the standard VaR tests. These levels for ES are approximately valid for a wide range of distributions, as evidenced by table 1.

The non-test criterion we use to compare ES forecasts is thus the rate of violation, defined as the proportion of observations for which the actual return is more extreme than the forecasted ES level. This rate, denoted ESRate, is:

$$\mathrm{ESRate} = \frac{\sum_{t=T+1}^{T+n} I(y_t < \mathrm{ES}_t)}{n},$$

where T is the estimation sample size and n is the forecast sample size. As above, a series of ES forecasts should have

Table 1. Nominal values and levels for ES for some common distributions.

					δ_α			
α	$N(0,1)$	AL	$t^*(10)$	$t^*(6)$	$t^*(4)$	$Sk-t^*(6)$	$Sk-t^*(4)$	TW
$\mathrm{ES}_{0.01}$	−2.6652	−3.6382	−3.0082	−3.2925	−3.6915	−3.7558	−4.2945	−3.474,−3.134
$\delta_{0.01}$	0.003847	0.003679	0.003601	0.003430	0.003212	0.003425	0.003208	0.0035,0.0037

Notes: AL refers to the Asymmetric Laplace distribution under the specification in Chen *et al*. (2012). TW refers to the two-sided Weibull distribution of Malevergne and Sornette (2004); δ_α is independent of the single parameter of the AL; the two parameters of the TW have been estimated from the data employed in this paper.

ESRate close to the nominal level δ_α. As also standard for VaR, the ratio ESRate/δ_α, called the ES ratio, is employed to compare competing ES forecasts: models with ES ratio ≈ 1 are most desirable. When ES ratio < 1, risk and loss forecasts are conservative (higher than actual), while alternatively, when ES ratio > 1, risk estimates are lower than actual and financial institutions may not allocate sufficient capital to cover likely future losses.

The UC, CC, independence, DQ and VQR tests, using either the exact value of δ_α for parametric GARCH models, or the approximate δ_αs of 0.36% quantile level for the other models, are also employed.

6. Empirical results

Six daily international stock market indices are analysed: the S&P 500 (US); FTSE 100 (UK); NASDAQ (US); DAX (Germany); SMI (Swiss) and the HANG SENG Index (Hong Kong, HK). Daily closing price index data from 6 April 2000 to 18 September 2014 are obtained from Thomson-Reuters Tick history, along with 1 min open, close, high and low prices for each day. The percentage log return series are generated as $y_t = (\ln(P_t) - \ln(P_{t-1})) \times 100$, where P_t is the closing price index or closing exchange rate on day t. Realized measures are obtained using the formulas in section 3. Market-specific non-trading days were removed from each series.

The full data period is divided into an estimation sample: 6 April 2000 to 31 December 2007, of $\approx T = 2000$ days; and a forecast sample: approximately $n = 1650$ trading days from 1 January 2008 to 18 September 2014. The latter period includes most, if not all, of the effects of the global financial crisis (GFC) on each market. Small differences in forecast sample sizes and end-dates occurred across markets, due to market-specific non-trading days. The exact in-sample sizes T and forecast sample sizes n are given in table 2. The value $q = 66$ is employed in the calculation of the realized measures RVSc and RRSc, which is equivalent to scaling using the last three months of trading days; this also explains why the estimation sample data starts in April 2000 and not January 2000. All series display the standard properties of daily asset returns: positive excess kurtosis, persistent heteroskedasticity and mostly mild, negative skewness.

For each series, the CARE-X-SAV and CARE-X-IG models are estimated using an MCMC burn-in sample size of 15 000 iterations, followed by a sampling period of 5000 iterations. To assess mixing and convergence, for each model the MCMC method is run from five different, randomly generated starting positions, for each market, at $\alpha = 0.01, 0.05$, with starting values for the parameters chosen to be widely varying and to lie on both sides of the estimated posterior means; convergence to the same posterior distribution is clear in all five runs for each parameter in each model, well before the end of the burn-in sample, in each market. Gelman's R statistics (see Gelman *et al*. 2005, p. 296) over these five runs are typically between 1.001 and 1.10 over all parameters; all highlighting fast mixing and clear and efficient convergence. Parameter estimates from all the models are not shown to save space.

6.1. Forecasting study

Expectiles, used to forecast VaR, and then ES levels are forecast 1 day ahead, for each day in the forecast sample of $m \approx 1650$ returns, in each series, using a range of competing CARE-X models, as well as non-parametric, semi-parametric and fully parametric competing specifications. For non-parametric methods, many financial institutions use 'historical simulation' to forecast VaR and ES; i.e. they employ sample percentiles for VaR estimation, and sample averages beyond those percentiles as ES forecasts. Two commonly used sample sizes are employed: the last 100 days (HS100) and the last 250 days (HS250, approximately one trading year). The two CARE-X models are considered, separately with each of the realized measures: RV, RVSc, RVSS, RR, RRSc, RRSS. The CARE-R-IG and SAV models of Gerlach and Chen (2016), each with Ra and RaO as inputs, as well as the CARE-IG and SAV models of Taylor (2008), with returns (Re) as input, are included. These models are each estimated in turn by the Bayesian MCMC method proposed. A range of popular GARCH specifications, including with Gaussian (GARCH-G), Student-*t* (GARCH-t) and Asymmetric Laplace (GARCH-AL; Chen *et al*. 2012) errors, are also considered. Further, a filtered GARCH (GARCH-HS) approach, where a GARCH-t is fit to the in-sample data, then a standardized VaR and ES are estimated via HS (using all the in-sample data) from the sample of returns (e.g. $y_1, \ldots, y_T$ divided by their GARCH-estimated conditional standard deviation (i.e. $y_t/\hat{\sigma}_t$). Then final forecasts of VaR, ES are found by multiplying the standardized VaR, ES estimates by the forecast $\hat{\sigma}_{t+1}$ from the GARCH-t model. All these GARCH models are estimated by ML, using the Econometrics toolbox in Matlab (GARCH-G, GARCH-t, GARCH-HS) or code developed by the authors (GARCH-AL). These models are fit by ML, since that is the standard choice in the literature; similar results to those reported below for these GARCH models were found by Chen *et al.* (2012) and Chen and Gerlach (2013) when estimating via Bayesian MCMC methods on similar data. The realized GARCH model of Hansen *et al*. (2011), with RV as an

Table 2. Counts of 1% VaR violations during the forecast period in each market.

Model	FTSE	SP500	HangSeng	Nasdaq	SMI	DAX	Avg.	Median
HS100	**28**	25	**27**	**27**	26	**30**	27.167	27
HS250	24	25	**25**	**26**	24	21	24.167	24.5
GARCH-HS	16	22	20	**29**	24	21	21.833	21
GARCH-G	**33**	**43**	**33**	**42**	**36**	**33**	**36.667**	**34.5**
GARCH-t	**26**	25	**26**	**32**	**27**	24	26.667	26
GARCH-AL	**7**	11	10	$\boxed{12}$	$\boxed{16}$	12	11.333	11.5
GJR-t	**31**	**27**	20	**30**	**31**	25	27.333	28.5
EGARCH-t	**32**	**30**	20	**30**	**30**	23	27.500	30
RG-RV-tG	20	**27**	**34**	**26**	18	**26**	25.167	26
RG-RR-tG	15	**28**	**30**	$\boxed{22}$	**31**	23	24.833	25.5
Re-IG	13	**26**	17	23	20	20	19.833	20
Re-SAV	20	**26**	$\boxed{16}$	24	22	21	21.500	21.5
Ra-IG	18	22	19	**27**	19	25	21.667	20.5
Ra-SAV	$\boxed{17}$	21	18	**26**	25	25	22.000	23
RaO-IG	18	20	20	**29**	20	24	21.833	20
RaO-SAV	$\boxed{17}$	$\boxed{19}$	21	**31**	24	23	22.500	22
RV-IG	19	23	**36**	**33**	22	25	26.333	24
RV-SAV	20	25	**36**	**28**	24	23	26.000	24.5
RVSS-IG	15	23	17	**30**	21	21	21.167	21
RVSS-SAV	15	**26**	$\boxed{16}$	**31**	22	23	22.167	22.5
RVSc-IG	18	**29**	19	**29**	21	**27**	23.833	24
RVSc-SAV	19	**26**	18	**32**	23	23	23.500	23
RR-IG	16	20	23	$\boxed{22}$	21	18	20.000	20.5
RR-SAV	16	21	14	**26**	19	$\boxed{17}$	18.833	$\boxed{18}$
RRSS-IG	11	22	$\boxed{16}$	$\boxed{22}$	21	16	$\boxed{18.000}$	18.5
RRSS-SAV	13	20	18	25	21	19	19.333	19.5
RRSc-IG	16	**26**	18	**29**	21	20	21.667	20.5
RRSc-SAV	$\boxed{17}$	**26**	$\boxed{16}$	**30**	19	19	21.167	19
FC-Min	**3**	**2**	**3**	**2**	**3**	**3**	**2.667**	**3**
FC-Max	**51**	**61**	**59**	**61**	**55**	**57**	**58.667**	**59**
FC-Mean	15	18	21	24	20	$\boxed{17}$	19.167	19
FC-Med	18	20	19	**29**	21	21	21.333	20.5
T	1943	1905	1890	1892	1930	1936		
n	1697	1676	1631	1672	1666	1692		

Notes: Boxes indicate the model closest to its nominal violation rate, bold indicates the violation rate is significantly different to 1% by the UC test, in each column. 'Avg.' is the average over the six series; 'FC-' stands for forecast combination.

input (RG-RV-tG), and the realized GARCH model of Gerlach and Wang (2016), with RR as an input (RG-RR-tG), both with Student-t observation errors and Gaussian measurement errors, are also included; estimated by the MCMC methods in Gerlach and Wang (2016). Finally, the forecast combination methods suggested by Chang *et al.* (2011) and McAleer *et al.* (2013) are included: the forecasts from all the methods on each day are combined via their mean (FC-Mean), median (FC-Med), minimum (FC-Min) and maximum (FC-Max), creating four extra VaR forecast series for inclusion.

For each day y_{T+t}, $t = 1, \ldots, n$, in the forecast sample, parameters are estimated for each model and method, employing the fixed window size T of data $(y_t, \ldots, y_{T+t-1})$ as observations (except HS100 and HS250 as above). Forecasts are then calculated for the next day's (time $T + t$) α-level VaR and ES. Under Bayesian estimation, each set of parameter iterates is substituted into the respective CARE-X equation to generate an expectile (and thus VaR) forecast, which is then transformed into an ES forecast via (2). Thus 5000 (post burn-in) MCMC ES forecast iterates are generated each day for each CARE model, which are then averaged to give the final posterior mean ES forecast. Each MCMC run takes from 5 to 10 s, estimating τ by grid search takes a similar amount of time, while ALS estimation takes less than 1 s, on a standard desktop PC.

6.2. *Forecasting VaR*

Table 2 shows the numbers of violations from the 1% VaR forecasts in each series, as well as the forecast and estimation sample sizes, plus the mean and median of these numbers across the markets, for each model. For series with $n \approx 1670$ we expect 16 or 17 violations from the 1% VaR forecasts on average. It is clear that the GARCH-G model consistently has the highest numbers of 1% VaR violations, and significantly under-estimates risk levels, in all series. The GARCH-AL always has the lowest number of violations in each series (excepting the FC-Min series) and clearly tends to over-estimate VaR levels, but not significantly (excepting the FTSE). The HS100, EGARCH-t, GJR-t and GARCH-t models significantly under-estimate risk levels in most series. It is also clear that CARE-X models with RR or RRSS as an input, are consistently and on average closest to the nominal violation rate expected, as also

Table 3. Counts of model rejections for six formal VaR forecast assessment tests, across the six markets.

$\alpha = 0.01$	IND	UC	CC	DQ1	DQ4	VQ	Total
HS100	1	**5**	**5**	5	**6**	2	**6**
HS250	4	2	4	5	**6**	1	**6**
GARCH-HS	$\boxed{0}$	1	1	1	2	1	3
GARCH-G	$\boxed{0}$	**6**	**6**	**6**	**6**	**5**	**6**
GARCH-t	$\boxed{0}$	4	2	2	5	2	**6**
GARCH-AL	$\boxed{0}$	1	1	$\boxed{0}$	3	3	5
GJR-t	$\boxed{0}$	4	4	5	5	3	5
EGARCH-t	$\boxed{0}$	4	4	5	4	$\boxed{0}$	5
RG-RV-tG	$\boxed{0}$	4	2	2	1	3	5
RG-RR-tG	$\boxed{0}$	3	3	3	2	3	4
Re-IG	$\boxed{0}$	$\boxed{0}$	1	$\boxed{0}$	2	$\boxed{0}$	3
Re-SAV	$\boxed{0}$	$\boxed{0}$	1	$\boxed{0}$	4	$\boxed{0}$	4
Ra-IG	$\boxed{0}$	1	1	$\boxed{0}$	3	1	3
Ra-SAV	$\boxed{0}$	1	1	1	4	1	4
RaO-IG	$\boxed{0}$	1	1	1	1	1	2
RaO-SAV	$\boxed{0}$	1	1	1	1	1	$\boxed{1}$
RV-IG	$\boxed{0}$	2	2	2	3	2	4
RV-SAV	$\boxed{0}$	2	2	2	2	1	3
RVSS-IG	1	1	1	2	3	2	4
RVSS-SAV	1	2	2	2	3	2	4
RVSc-IG	$\boxed{0}$	3	3	2	4	$\boxed{0}$	4
RVSc-SAV	1	2	2	2	5	$\boxed{0}$	5
RR-IG	1	$\boxed{0}$	$\boxed{0}$	2	3	$\boxed{0}$	3
RR-SAV	1	$\boxed{0}$	1	1	3	1	3
RRSS-IG	1	$\boxed{0}$	$\boxed{0}$	1	3	3	3
RRSS-SAV	1	$\boxed{0}$	$\boxed{0}$	1	3	1	3
RRSc-IG	1	2	1	2	4	1	4
RRSc-SAV	1	2	1	2	3	1	4
FC-Min	$\boxed{0}$	**6**	**6**	**6**	2	**6**	**6**
FC-Max	1	**6**	**6**	**6**	**6**	**6**	**6**
FC-Mean	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	2	$\boxed{0}$	2
FC-Med	$\boxed{0}$	1	1	1	2	$\boxed{0}$	2

Notes: Boxes indicate the favoured model, bold indicates the least favoured model, in each column. 'Total' indicates the number of markets in which the model was rejected by at least one test.

are the GARCH-HS and forecast combined series that is the mean (FC-Mean) of all VaR forecasts each day. All models consistently under-estimate risk levels, with the most accurate producing marginally too many violations on average, while all the CARE and CARE-X (except the CARE-X-RV models) are more accurate than the other models considered (except the GARCH-HS).

Each model's VaR forecasts are formally tested for accuracy and independence, across the six series. Table 3 shows counts of the number of rejections, at the 5% significance level, for each model/method, over the six return series, for each of the tests considered: UC, independence, CC, DQ (1 and 4 lags) and VQ tests, as well as a total (for rejection by at least 1 test) across all 6 markets. For 1% VaR forecasting, clearly the HS100, HS250 and all parametric GARCH models are the worst performing, being rejected in 5 or 6 (i.e. all) markets by at least one test; clearly these are highly inadequate methods for 1% VaR forecasting in these series. Though the GARCH-AL is rejected only a maximum of 3 times by any single test, across markets it is rejected 5 times, as is the RG-RV-tG model. All the CARE and CARE-X models do better regarding minimizing the number of test rejections across markets. The CARE-RaO models are only rejected in 1 and 2 series, respectively, whilst the CARE-X models with RR and RRSS, and the GARCH-HS, are rejected in 3 of the series; the forecast combination series 'FC-Mean' and 'FC-Med' are rejected in only 2 markets.

Clearly, for 1% VaR forecasting, the CARE-X models are the most favoured and most accurate in terms of the criteria applied; the FC-Mean and FC-Med series are also quite accurate. All other models/methods perform less favourably and can be rejected as adequate 1% VaR forecast methods for these series and the time period chosen, with the exception of the GARCH-HS method. Clearly, GARCH models with Gaussian or Student-*t* errors are simply inadequate at accurately modelling the tail in the series considered over the forecast period: their estimated VaRs tend to consistently under-estimate VaR levels, leading to significantly too many violations. In contrast, the tails of the AL distribution are too fat, and tend to over-estimate the ES levels. The HS methods simply cannot track either the levels or dynamics of VaR, not surprising from such an ad hoc method. Only the CARE-X models, especially those with RR and RRSS as input, could consistently and reasonably accurately forecast VaR levels and dynamics among the models considered; not being constrained by a parametric distributional assumption (e.g. Student-*t* errors) is apparently a clear advantage of these models. The extra gains from including the information on the RR, or intra-day range plus overnight price movements, via the CARE-X models, were fairly clear in the 1% VaR case.

Table 4 shows summary statistics for the VaR violation rates across the six series for each model, divided by 0.01 and denoted as VRate ratio; the ideal is VRate ration = 1. These

Table 4. Forecast comparisons of VRate ratio $= \hat{\alpha}/0.01$: $\alpha = 0.01$.

Model	Mean	Median	RMSE
HS100	1.6243	1.6324	0.6303
HS250	1.4459	1.4661	0.4578
GARCH-HS	1.3065	1.2694	0.3913
GARCH-G	**2.1928**	**2.0921**	**1.2198**
GARCH-t	1.5951	1.5631	0.6155
GJR-t	1.6328	1.7026	0.6718
EGARCH-t	1.6427	1.7921	0.6905
GARCH-AL	0.6782	0.6828	0.3603
RG-RV-tG	1.5077	1.5458	0.6033
RG-RR-tG	1.4883	1.5150	0.5969
Re-IG	1.1863	1.1913	0.3095
Re-SAV	1.2847	1.2808	0.3383
Ra-IG	1.2952	1.2388	0.3546
Ra-SAV	1.3153	1.3653	0.3789
RaO-IG	1.3056	1.2134	0.3757
RaO-SAV	1.3462	1.3234	0.4384
RV-IG	1.5785	1.4249	0.6939
RV-SAV	1.5587	1.4661	0.6467
RVSS-IG	1.2657	1.2508	0.3896
RVSS-SAV	1.3250	1.3399	0.4617
RVSc-IG	1.4244	1.4281	0.5044
RVSc-SAV	1.4047	1.3699	0.4896
RR-IG	1.1977	1.2269	$\boxed{0.2518}$
RR-SAV	1.1257	$\boxed{1.0726}$	0.2630
RRSS-IG	$\boxed{1.0773}$	1.1207	0.2557
RRSS-SAV	1.1569	1.1581	0.2680
RRSc-IG	1.2958	1.2213	0.3997
RRSc-SAV	1.2653	1.1317	0.4021
FC-Min	**0.1595**	**0.1770**	0.8410
FC-Max	**3.5087**	**3.5471**	**2.5124**
FC-Mean	1.1477	1.1372	$\boxed{0.2350}$
FC-Med	1.2758	1.2172	0.3497

Notes: Boxes indicate the favoured model, bold indicates a model rejected by the UC test, in each column. 'RMSE' stands for square root of the average squared difference between the six ratios and their expected value of 1.

summary statistics confirm that discussed above: all the GARCH models are anti-conservative, except for the GARCH-AL and GARCH-HS, the former of which has ≈ 68% of the VaR violations expected and clearly over-estimates VaR levels. The HS100 and HS250 models are also highly anti-conservative, the best of these with around 45% more VaR violations than expected. All the CARE-X models have VRates closer to 1 than all the other models, except the CARE-X-RV and GARCH-HS models. The CARE-X models with VRates closest to 1 in average, median and RMSE are the CARE-X models that employ RR and RRSS as inputs.

In summary, the GARCH, Realized GARCH and HS models are all clearly and relatively inaccurate for 1% VaR forecasting in these markets. The CARE-type models are more accurate at 1% VaR forecasting and are generally rejected in fewer markets, with those employing RR and RRSS being the most accurate in terms of violation rates. However, the CARE-X employing RaO was rejected in the fewest markets and also did well on most measures. These were the best performing group of models. There does seem to be an advantage, especially in terms of lower RMSE, in taking the Mean VaR forecast (FC-Mean) combination across these models/methods.

6.3. *Forecasting ES*

Figure 1 shows the returns in the forecast period for the US market, as well as the 1% ES forecasts for three different methods: GARCH-G, GARCH-t and CARE-RR-SAV. The forecasts for the GARCH with Gaussian errors (GARCH-G) are often the least extreme for each day, as expected, while those for the GARCH with Student-*t* (GARCH-t) distributed errors are usually, and clearly, the most extreme each day, excepting the start of the sample and during the crisis period, where the CARE-RR-SAV is often the most extreme. In fact, the GARCH-t gives the most extreme ES forecast on 1170 out of 1676 days in the forecast sample and is more extreme than the GARCH-G forecast for every day in the forecast sample; for the remaining 506 days the CARE-RR-SAV gives the most extreme ES forecast. However, the CARE-RR-SAV model forecasts are very similar to those from the GARCH-G model on many days, especially during low volatility periods; in fact for 376 days in the forecast sample the CARE-RR-SAV's ES forecast is less extreme than both the GARCH-G and GARCH-t forecasts. This is remarkable given that the GARCH-G had 23 ES violations, the GARCH-t had 11 ES

Table 5. VaR and ES joint loss function values, using equation (25), across the markets; $\alpha = 0.01$.

Model	FTSE	SP500	HangSeng	Nasdaq	SMI	DAX	Mean	Rank
HS100	1749.0	1773.9	1733.0	1810.4	1734.0	1847.7	1774.7	30.17
HS250	1770.1	1786.7	1738.0	1808.8	1756.4	1825.6	1780.9	30.67
GARCH-HS	1686.1	1686.4	1689.4	1709.3	1672.5	1767.4	1701.8	21.33
GARCH-G	1717.0	1742.9	1714.0	1746.8	1746.6	1808.4	1745.9	28.67
GARCH-t	1694.6	1694.6	1693.8	1717.5	1696.7	1775.5	1712.1	24.50
GJR-t	1711.1	1687.4	1684.6	1708.6	1693.4	1782.2	1711.2	22.83
EG-t	1744.7	1723.4	1687.9	1737.1	1686.7	1786.7	1727.8	25.50
GARCH-AL	1698.0	1689.7	1703.0	1711.1	1672.6	1761.3	1705.9	22.83
RG-RV-tG	1666.9	1655.4	1707.7	1694.3	1641.3	1726.7	1682.1	14.83
RG-RR-tG	1660.6	1639.3	1688.0	$\boxed{1678.5}$	1641.1	$\boxed{1711.0}$	1669.7	6.17
Re-IG	1697.4	1682.0	1684.3	1716.0	1679.6	1764.6	1704.0	21.50
Re-SAV	1695.0	1728.2	1679.1	1728.1	1687.0	1751.8	1711.5	22.33
Ra-IG	1661.8	1650.5	1682.5	1695.6	1679.1	1735.8	1684.2	15.00
Ra-SAV	1656.1	1645.3	1677.6	1699.2	1687.4	1735.0	1683.4	12.83
RaO-IG	1661.8	1651.1	1672.2	1698.5	1660.1	1738.4	1680.3	13.33
RaO-SAV	1656.2	1646.4	1670.6	1700.1	1661.9	1732.8	1678.0	10.67
RV-IG	1665.7	1649.0	1719.6	1708.2	1644.0	1730.7	1686.2	16.67
RV-SAV	1667.6	1646.4	1712.0	1705.2	1648.2	1736.0	1685.9	17.33
RVSc-IG	1669.2	1654.2	1686.5	1709.9	1642.0	1738.6	1683.4	17.83
RVSc-SAV	1666.0	1650.1	1684.1	1713.1	1652.2	1733.4	1683.1	16.33
RVSS-IG	1654.8	1643.9	1674.3	1689.3	$\boxed{1637.4}$	1719.2	1669.8	$\boxed{4.83}$
RVSS-SAV	$\boxed{1654.6}$	1643.2	$\boxed{1670.1}$	1690.7	1641.5	1731.2	1671.9	5.83
RR-IG	1663.6	1641.3	1689.0	1686.7	1643.7	1713.5	1673.0	9.67
RR-SAV	1664.0	1639.8	1686.1	1685.4	1642.0	1717.5	1672.5	8.17
RRSc-IG	1668.3	1651.9	1674.8	1692.9	1640.7	1718.8	1674.6	10.33
RRSc-SAV	1665.5	1648.4	1673.1	1695.2	1639.7	1724.6	1674.4	9.00
RRSS-IG	1659.0	1640.6	1678.0	1684.9	1642.2	1711.4	$\boxed{1669.4}$	6.17
RRSS-SAV	1655.7	$\boxed{1637.6}$	1678.0	1684.9	1643.4	1723.6	1670.5	5.67
FC-Min	1726.5	1722.1	1713.0	1742.5	1704.3	1776.4	1730.8	28.00
FC-Max	**1842.3**	**1858.9**	**1802.1**	**1868.8**	**1851.0**	**1904.4**	**1854.6**	**32.00**
FC-Mean	1661.3	1653.9	1672.8	1691.6	1645.1	1727.6	1675.4	11.00
FC-Med	1658.9	1647.5	1671.1	1691.0	1641.0	1726.3	1672.6	6.83

Notes: Boxes indicate the favoured model, blue shading indicates the 2nd and 3rd ranked models, bold indicates the least favoured, red shading indicates the 2nd and 3rd lowest ranked models, in each column. 'Rank' is the average rank across the six markets for the loss function, over the 32 models: lower is better.

violations and the CARE-RR-SAV had only 7 ES violations, which is very close to the nominal rate expected. These patterns for and among the forecasts for the different methods are fairly consistent across the six return series, for 1% ES forecasting. The patterns, combined with the subsequently discussed accuracy results, may indicate the superior efficiency of the RR and RV inputs when included in the CARE model, compared to traditional GARCH models. In particular, these patterns show that lower capital allocations (which correspond to less extreme ES forecasts) can be achieved on many days, whilst simultaneously improving the accuracy of ES forecasts, when employing CARE-X models.

Figure 2 shows the average standardized ES residual, where these are standardized by the VaR forecasts, in each of the six return series, for each of the models/methods, from the 1% ES forecasts. Since there is no practical difference between these figures for SAV or IG CARE-type models, only the CARE-SAV-type model results are shown. These averages should (theoretically) be zero for an accurate ES forecast method. Each symbol is the average ES residual for a particular return series, over the models/methods, while each large black circle shows the mean of the six ES residual averages for each model/method; a reference line is at zero. Clearly, the GARCH-G (G-G), HS100 and HS250 methods consistently under-predict ES levels across the six series, since all their average ES residuals are negative. This indicates that over the full forecast period these three methods are typically anti-conservative in ES forecasting; while the GARCH-AL (G-AL) and RG-RR-tG models consistently over-predict ES levels and are the only consistently conservative methods. The remaining models, including the GARCH models with HS (G-HS) and Student-*t* errors (G-t), the asymmetric GARCH models, the RG-RV-tG model and all the CARE-X and CARE models have average ES residuals reasonably close to, and on both sides of, 0. It is hard to separate this latter group of models based on their standardized ES residuals.

Table 5 shows the loss function values S, calculated using equation (25), which jointly assess the accuracy of each model's VaR and ES series, during the forecast period for each market. On this measure, the CARE-X models using RRSS and RVSS do best overall, having lower loss than most other models in most markets and being consistently ranked higher on that measure; the RG-RR-tG model also does well on this criterion. The ad hoc HS methods do the worst, across the six markets, almost always ranking in the bottom two places across the individual models, only trailed by the forecast combination method FC-Max. Generally the CARE-X models with RRSS and RVSS and the RG-RR-tG model are more highly ranked, having lower loss, than all other models in most markets. These models, together with the CARE-RR and CARE-RRSc

Table 6. Counts of ES violations during the forecast period in each market, $\alpha = 0.01$.

Model	FTSE	SP500	HangSeng	Nasdaq	SMI	DAX	Mean	Median
HS100	**22**	**20**	**22**	**21**	**21**	**21**	21.167	21
HS250	**15**	**15**	**15**	**16**	**16**	10	14.500	15
GARCH-HS	8	9	8	8	9	8	8.333	8
GARCH-G	**17**	**23**	**18**	**23**	**24**	**17**	20.333	20.5
GARCH-t	**13**	11	8	8	**16**	9	10.833	10
GJR-t	11	10	7	10	**17**	**12**	11.167	10.5
EGARCH-t	10	11	7	11	**18**	10	11.167	10.5
GARCH-AL	**1**	4	3	4	3	5	3.333	3.5
RG-RV-tG	7	$\boxed{6}$	15	9	6	7	8.333	7
RG-RR-tG	7	5	10	5	11	8	7.667	7.5
Re-IG	9	5	8	10	9	7	8.000	8.5
Re-SAV	9	$\boxed{6}$	8	11	10	5	8.167	8.5
Ra-IG	$\boxed{6}$	4	9	9	9	**13**	8.333	9
Ra-SAV	5	5	9	$\boxed{6}$	**12**	10	7.833	7.5
RaO-IG	$\boxed{6}$	4	$\boxed{6}$	$\boxed{6}$	7	10	6.500	$\boxed{6}$
RaO-SAV	5	5	7	7	10	9	7.167	7
RV-IG	$\boxed{6}$	7	**13**	8	8	$\boxed{6}$	8.000	7.5
RV-SAV	7	9	**15**	8	8	5	8.667	8
RVSS-IG	5	8	8	7	7	7	7.000	7
RVSS-SAV	7	8	9	8	9	$\boxed{6}$	7.833	8
RVSc-IG	7	7	9	8	7	9	7.833	7.5
RVSc-SAV	7	$\boxed{6}$	9	8	9	$\boxed{6}$	7.500	7.5
RR-IG	4	5	11	7	11	$\boxed{6}$	7.333	6.5
RR-SAV	4	7	8	7	5	5	$\boxed{6.000}$	$\boxed{6}$
RRSS-IG	4	5	8	7	9	5	6.333	$\boxed{6}$
RRSS-SAV	4	$\boxed{6}$	7	$\boxed{6}$	9	5	6.167	$\boxed{6}$
RRSc-IG	9	$\boxed{6}$	7	8	10	7	7.833	7.5
RRSc-SAV	8	8	7	7	$\boxed{6}$	5	6.833	7
FC-Min	**1**	**1**	2	**1**	3	2	0.833	1
FC-Max	**49**	**34**	**43**	**37**	**36**	**37**	34.667	34.5
FC-Mean	5	5	7	7	$\boxed{6}$	3	5.500	5.5
FC-Med	7	5	7	$\boxed{6}$	$\boxed{6}$	$\boxed{6}$	6.167	$\boxed{6}$
T	1943	1905	1890	1892	1930	1936		
n	1697	1676	1631	1672	1666	1692		

Notes: Boxes indicate the model closest to its nominal violation rate, bold indicates the violation rate is significantly different to nominal by the UC test, in each column.

models, consistently outperform the all other models, as well as the forecast combination methods FC-Mean and FC-Median in most markets.

Table 6 shows the numbers of violations from the 1% ES forecasts in each series, as well as the forecast and estimation sample sizes. For series with $n \sim 1650$ we expect ≈ 6 violations from the 1% ES forecasts from a violation rate of ≈ 0.0036. In all six series at least one CARE-X model achieves this, a standard CARE model achieves it in only one series, and no other model achieves this in any series. The FC-Mean and FC-Med forecast combination series also achieve this in 1 and 3 series, respectively. It is clear that the GARCH-G, HS100, HS250 consistently have the highest numbers of 1% ES violations in all series and significantly under-estimate 1% ES risk levels. The GARCH-AL always has the lowest number of violations in each series (excepting the FC-Min series) and clearly tends to over-estimate ES levels, but not significantly (excepting the FTSE). The GARCH-t is clearly better the GARCH-G, but still consistently has too many violations, under-estimating risk levels, as do the EGARCH-t and GJR-t models, in all series. It is clear that CARE-X models with RR, RRSS or RaO as input, are consistently close to the 6 violations expected for accurate ES forecast models. The combined series FC-Mean and FC-Med are also consistently close to the nominal rate. CARE-X models with RV, RVSc, RVSS as input consistently under-estimate risk levels, with marginally too many violations on average, but seem more accurate than the other anti-conservative models discussed so far.

Figure 3 shows the ESRate ratios for the 1% ES forecasts for each model in the forecast period over the six series. Each symbol is the ESRate divided by its expected nominal rate, for a particular return series over the models/methods (marked on *x*-axis), while each larger black circle represents the average ES violation rate ratio across all 6 markets. These ratios should be close to 1 for each model. The dashed lines mark the points where the p-value from the UC test is closest to 0.05; models with ratios above the top dashed line, or below the bottom dashed line, are rejected by the UC test. Again, the GARCH-G (G-G), HS100 and HS250 models have ESRates well above nominal: these models are all rejected by the UC test in all or most markets. Further, to a lesser extent the GARCH-HS (G-HS), GARCH-t, GJR-t and EGARCH-t (EG-t) models are also typically anti-conservative in their ES risk forecasts, having

Table 7. Counts of model rejections for seven formal ES forecast assessment tests, across the six markets.

$\alpha = 0.01$	boot.	IND	UC	CC	DQ1	DQ4	VQ	Total
HS100	**6**	2	**6**	**6**	**6**	**6**	2	**6**
HS250	3	2	5	5	**6**	**6**	2	**6**
GARCH-HS	1	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	4	$\boxed{0}$	5
GARCH-G	5	$\boxed{0}$	**6**	**6**	**6**	**6**	3	**6**
GARCH-t	1	$\boxed{0}$	2	2	2	5	1	5
GJR-t	1	$\boxed{0}$	2	2	2	5	1	5
EG-t	$\boxed{0}$	$\boxed{0}$	1	1	2	4	1	5
GARCH-AL	1	$\boxed{0}$	1	1	$\boxed{0}$	3	1	5
RG-RV-tG	1	$\boxed{0}$	1	1	1	2	2	3
RG-RR-tG	2	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	1	2	1	3
Re-IG	1	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	5	$\boxed{0}$	5
Re-SAV	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	4	$\boxed{0}$	4
Ra-IG	$\boxed{0}$	$\boxed{0}$	1	1	1	3	$\boxed{0}$	3
Ra-SAV	$\boxed{0}$	$\boxed{0}$	1	$\boxed{0}$	$\boxed{0}$	2	$\boxed{0}$	2
RaO-IG	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{1}$	$\boxed{0}$	$\boxed{1}$
RaO-SAV	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	3	$\boxed{0}$	3
RV-IG	$\boxed{0}$	$\boxed{0}$	1	1	1	3	1	3
RV-SAV	$\boxed{0}$	$\boxed{0}$	1	1	1	3	1	3
RVSc-IG	1	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	4	$\boxed{0}$	4
RVSc-SAV	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	4	$\boxed{0}$	4
RVSS-IG	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	1	3	$\boxed{0}$	3
RVSS-SAV	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	2	$\boxed{0}$	2
RR-IG	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	1	3	1	3
RR-SAV	1	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	2	$\boxed{0}$	2
RRSc-IG	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	3	1	3
RRSc-SAV	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	1	4	$\boxed{0}$	4
RRSS-IG	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	1	3	1	3
RRSS-SAV	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	3	$\boxed{0}$	3
FC-Min	**6**	$\boxed{0}$	5	5	$\boxed{0}$	$\boxed{0}$	**6**	**6**
FC-Max	**6**	$\boxed{0}$	**6**	**6**	**6**	**6**	2	**6**
FC-Mean	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	4	1	4
FC-Med	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	3	1	3

Notes: Boxes indicate the favoured model, bold indicates the least favoured model, in each column. 'Total' indicates the number of markets in which the model was rejected by at least one test.

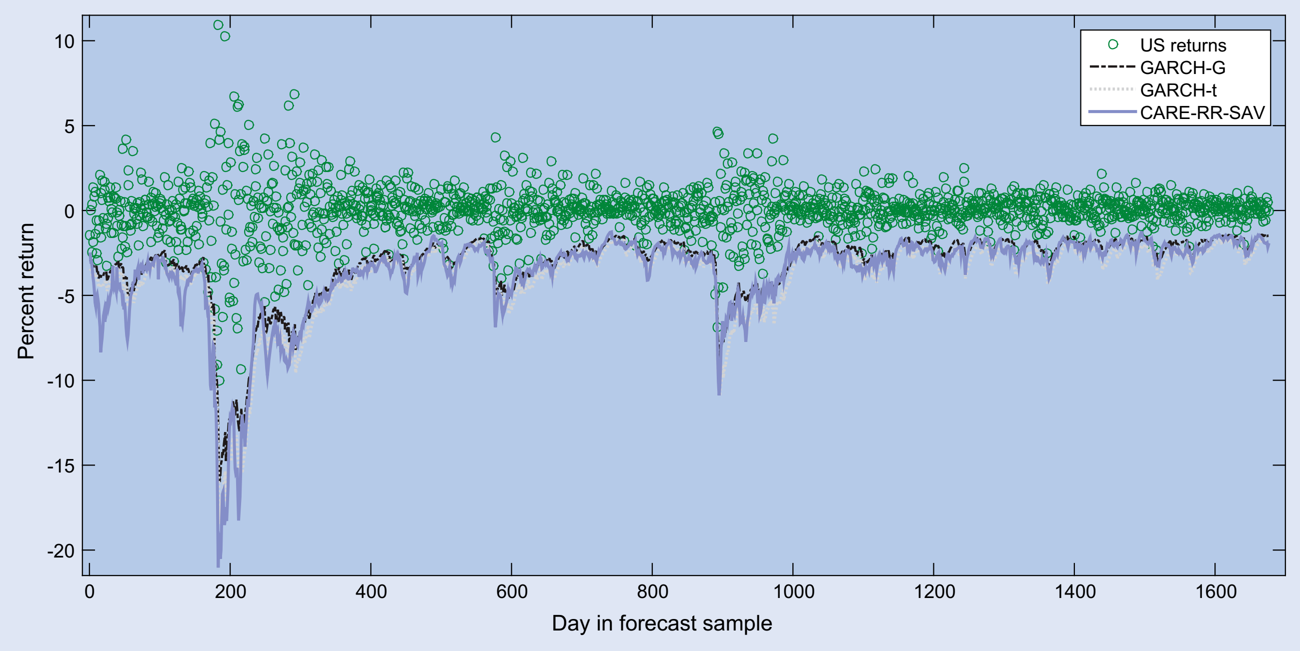

Figure 1. Some 1% ES forecasts and the US S&P500 forecast sample returns.

Table 8. Forecast comparisons of $\hat{\delta}/\delta$: $\alpha = 0.01$.

Model	FTSE	SP500	HangSeng	Nasdaq	SMI	DAX	Mean	Median	RMSE
HS100	**3.601**	**3.315**	**3.747**	**3.489**	**3.501**	**3.448**	**3.517**	**3.495**	**2.520**
HS250	**2.455**	**2.486**	**2.555**	**2.658**	**2.668**	1.642	2.411	2.520	1.454
GARCH-HS	1.310	1.492	1.362	1.329	1.501	1.313	1.384	1.346	0.393
GARCH-G	**2.604**	**3.567**	**2.869**	**3.576**	**3.745**	**2.612**	3.162	3.218	2.214
GARCH-t	**2.203**	1.799	1.362	1.320	**2.634**	1.466	1.797	1.632	0.931
GJR-t	1.864	1.635	1.192	1.650	**2.799**	**1.955**	1.849	1.757	0.979
EG-t	1.695	1.799	1.192	1.815	**2.963**	1.629	1.849	1.747	1.006
GARCH-AL	**0.159**	0.645	0.497	0.647	0.487	0.799	0.539	0.571	0.502
RG-RV-tG	1.146	$\boxed{0.994}$	**2.555**	1.495	$\boxed{1.000}$	1.149	1.390	1.147	0.671
RG-RR-tG	1.146	0.829	1.703	0.831	1.834	1.313	1.276	1.230	0.477
Re-IG	1.473	0.829	1.362	1.661	1.501	1.149	1.329	1.418	0.427
Re-SAV	1.473	$\boxed{0.994}$	1.362	1.827	1.667	0.821	1.358	1.418	0.503
Ra-IG	$\boxed{0.982}$	0.663	1.533	1.495	1.501	**2.134**	1.385	1.498	0.603
Ra-SAV	0.818	0.829	1.533	$\boxed{0.997}$	**2.001**	1.642	1.303	1.265	0.541
RaO-IG	$\boxed{0.982}$	0.663	$\boxed{1.022}$	$\boxed{0.997}$	1.167	1.642	1.079	1.009	0.304
RaO-SAV	0.818	0.829	1.192	1.163	1.667	1.477	1.191	1.178	0.365
RV-IG	$\boxed{0.982}$	1.160	2.214	1.329	1.334	$\boxed{0.985}$	1.334	1.245	0.535
RV-SAV	1.146	1.492	**2.555**	1.329	1.334	0.821	1.446	1.332	0.699
RVSc-IG	1.146	1.160	1.533	1.329	1.167	1.477	1.302	1.248	0.340
RVSc-SAV	1.146	$\boxed{0.994}$	1.533	1.329	1.501	$\boxed{0.985}$	1.248	1.237	0.333
RVSS-IG	0.818	1.326	1.362	1.163	1.167	1.149	1.164	1.165	0.241
RVSS-SAV	1.146	1.326	1.533	1.329	1.501	$\boxed{0.985}$	1.303	1.327	0.358
RR-IG	0.655	0.829	1.873	1.163	1.834	$\boxed{0.985}$	1.223	1.074	0.522
RR-SAV	0.655	1.160	1.362	1.163	0.834	0.821	$\boxed{0.999}$	$\boxed{0.997}$	0.246
RRSc-IG	1.473	0.994	1.192	1.329	1.667	1.149	1.301	1.261	0.373
RRSc-SAV	1.309	1.326	1.192	1.163	$\boxed{1.000}$	0.821	1.135	1.178	$\boxed{0.223}$
RRSS-IG	0.655	0.829	1.362	1.163	1.501	0.821	1.055	0.996	0.313
RRSS-SAV	0.655	$\boxed{0.994}$	1.192	$\boxed{0.997}$	1.501	0.821	1.027	0.996	0.270
FC-Min	**0.000**	**0.166**	**0.170**	**0.166**	**0.000**	0.328	0.138	0.166	0.869
FC-Max	**5.565**	**5.801**	**5.280**	**6.313**	**6.503**	**5.910**	**5.895**	**5.856**	**4.913**
FC-Mean	0.818	0.829	1.192	$\boxed{0.997}$	$\boxed{1.000}$	0.493	0.916	0.915	0.253
FC-Med	1.146	0.829	1.192	$\boxed{0.997}$	$\boxed{1.000}$	$\boxed{0.985}$	1.025	$\boxed{0.999}$	$\boxed{0.121}$

Notes: Boxes indicate the favoured model, bold indicates a model rejected by the UC test, in each column; blue shading indicates models ranked 2nd and 3rd highest for that column. 'RMSE' stands for square root of the average squared difference between the six ratios and their expected value of 1.

too many ES violations in all markets. The G-AL model is the only consistently conservative model, with ES violation rates well below that expected under the AL distribution, and hence this model over-estimates ES risk levels, though mostly not significantly. The CARE-X and CARE models all have ESRates closer to nominal on average, many with no rejections by the UC test. CARE-X models with RV, RVSS, RVSc, Ra and RaO as inputs seem to be marginally less accurate than the CARE-X models employing RR, RRSS and RRSc as inputs. Clearly the CARE-X models with RR, RRSS, RRSc as inputs are closest to nominal on average, and also the least variable across the series.

Each model's ES forecasts are formally tested for accuracy and independence of violations, across the six series. Table 7 counts the number of rejections, at the 5% significance level, for each model/method, over the six return series, for each of the tests considered: bootstrap t-test, UC, independence, CC, DQ (1 and 4 lags) and VQ tests, as well as a total across all 6 markets. For 1% ES forecasting, clearly the HS100, GARCH-G and HS250 are the worst performing models, being rejected in at least 5 markets by most tests, and rejected in all markets by at least one test: clearly these are highly inadequate methods for 1% ES forecasting in these series. Though the GARCH-AL is rejected only a maximum of 3 times by any single test, across markets it is rejected 5 times, as are the GARCH-t, GJR-t and EGARCH-t models. The CARE and CARE-X models do better than these regarding minimizing the number of rejections across markets. The CARE-RaO-IG model is only rejected in 1 series, whilst the CARE-RR-SAV, CARE-Ra-SAV and CARE-RVSS-SAV are rejected in only 2 series. Models with RR, RRSS, RRSc as inputs are generally rejected in 3 of the series, as is the forecast combination series FC-Med.

Clearly, for 1% ES forecasting, the CARE-X models are generally the most favoured and most accurate, and are among those with the lowest joint VaR and ES loss function values, especially those involving RR, RRSS, RVSS or RaO as input, in terms of the criteria applied. On the contrary, all other models/methods perform less favourably on most criteria, and can be rejected as adequate 1% ES forecast methods for these series and the time period chosen. Clearly, GARCH, GJR and EGARCH models are simply inadequate at accurately modelling the conditional distribution in the series considered over the forecast period; their tails are not fat enough and tend to consistently under-estimate ES levels, leading to too many violations in most series. In contrast, the tails of the AL distribution are too fat, and tend to over-estimate the ES levels. The HS methods simply cannot track either the levels or dynamics of ES, not surprising from such an ad hoc method. Only the CARE-X models could consistently and accurately forecast ES levels and dynamics of the models considered. Not

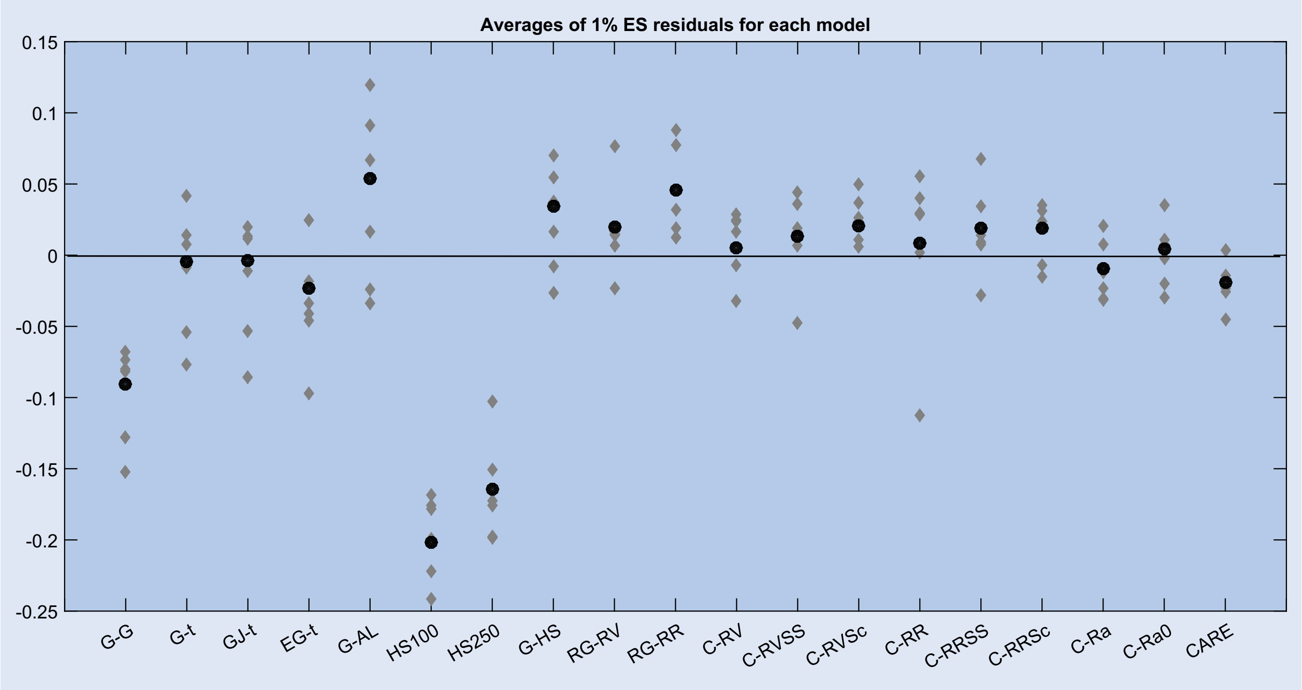

Figure 2. Averages of the standardized ES residuals $(r_t - \mathrm{ES}_t)/\mathrm{VaR}_t$, for 1% ES forecasts and returns that violate the 1% VaR forecast, for each of the six return series and each model (marked on x-axis). A reference line is at 0, where accurate models are expected to have their average ES residual; the large black circles indicate the mean of the six ES residual averages for each model.

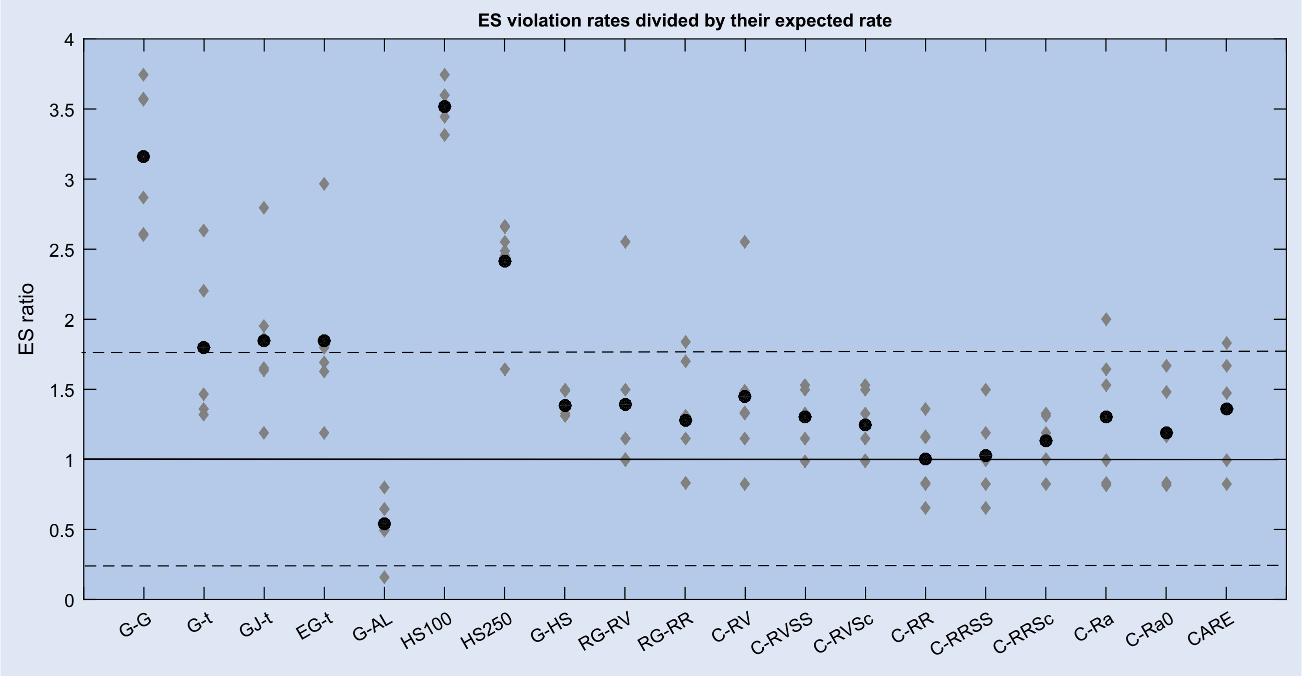

Figure 3. 1% ES violation rates, divided by the nominal rate (e.g. 0.0036), across all six series for each model considered. A large black circle indicates the mean of the six ESRates for each model. Three reference lines are drawn: the solid middle is at 1 (the target ratio); the two dashed lines contain the non-rejection region for the UC test (models with ESRate ratios above or below these lines are rejected).

being constrained by a parametric distributional assumption (e.g. Student-t errors) is apparently a clear advantage of these models. The extra gains from including the information on the RR, or intra-day range plus overnight price movements, via the CARE-X models, were fairly clear in the 1% ES case. The forecast combined series FC-Med also performed well under all measures.

Table 8 shows the ES violation rate ratios and summary statistics across the six series, taking the observed violation rates and dividing them by the expected rates: 0.0036 for 1%

ES for CARE, GARCH-type models with Student-*t* errors, RG-tG and HS methods, 0.0038 for Gaussian and 0.0037 for the GARCH-AL model. These summary statistics confirm the conclusions above: all the GARCH models are anti-conservative, except for the GARCH-AL, which has only 50% of the ES violations expected and clearly over-estimates the ES levels. All the CARE-X models have ratios closer to 1 than all the other models. The CARE-X models with ratios closest to 1 on average (and median) are the CARE-RR-SAV, RRSS-SAV, RRSS-IG, RaO-IG and RRSc-SAV models, all with average ratios within about 15% of the desired 1 and with the lowest RMSE from 1. The GARCH-G, GARCH-t, GJR-t, EGARCH-t,

HS100 and HS250 models are highly anti-conservative, the best of these with around 80% more ES violations than expected; the worst having 2.4 to 3.5 times too many ES violations. CARE-X models with RV, RVSS, RVSc and Ra as an input have between 25 and 45% too many violations, mirrored by the CARE models with returns as an input.

In summary, the parametric GARCH models, HS100 and HS250 models are all clearly anti-conservative at ES forecasting, at the 1% levels; excepting the GARCH-AL model that is quite conservative, clearly over-estimating ES levels at this risk level. All the CARE-X are more accurate than these, in terms of ES violation rates, with average and median ESRate ratio much closer to 1, and with much lower RMSE of the six ratios compared to 1. The CARE-type models were rejected in the least number of markets (1–3) compared to the other models. The CARE-RaO-IG was rejected in the fewest markets, whilst the CARE-X models with RR, RRSS had closest to nominal ES violation rates and CARE-X models with RRSS and RVSS had the lowest joint loss values, using equation (25). These were the best performing group of models. These individual models all out-performed the best forecast combination method, being FC-Med, i.e. taking the median of the ES forecasts across the models for each day, on all criteria considered, for 1% ES forecasting.

7. Conclusion

This paper considers dynamic expectile, Value at Risk and ES modelling and forecasting, incorporating information from a range of realized measures. The class of CARE-X models using these realized measures as inputs is proposed and developed. A psuedo-likelihood formulation is employed, allowing for a Bayesian MCMC method to be employed to estimate the proposed model class. An ES forecasting study, using a forecast sample during and after the GFC, reveals that the proposed CARE-X models are highly competitive, in terms of ES residuals, joint VaR and ES loss function values, VaR and ES violation rates and independence of VaR and ES violations, compared to a range of well-known models and methods, including standard CARE models, GARCH, Realized GARCH and historical simulation, across six return series over a seven-year forecast period. For ES forecasting, the best CARE-X model was highly competitive, and mostly better, than the best forecast combination series, for each criterion applied, and usually involved the realized measures based on the realized range: RRSS or RR, and/or the range plus overnight return: RaO.

Acknowledgements

The authors thank the Editors, Managing Editor and a referee for insightful comments that helped improve the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Richard Gerlach http://orcid.org/0000-0002-5656-4556

References

Acerbi, C. and Szekeley, B., Backtesting expected shortfall. *Risk*. December, 2014.

Aigner, D.J., Amemiya, T. and Poirier, D.J., On the estimation of production Frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *Int. Econ. Rev.*, 1976, **17**, 377–396.

Alizadeh, S., Brandt, M.W. and Diebold, F.X., Range-based estimation of stochastic volatility models or exchange rate dynamics are more interesting than you think. *J. Finance*, 2002, **57**, 1047–1092.

Andersen, T. and Bollerslev, T., Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *Int. Econ. Rev.*, 1998, **39**, 885–905.

Artzner, P., Delbaen, F., Eber, J.M. and Heath, D., Thinking coherently. *Risk*, 1997, **10**, 68–71.

Artzner, P., Delbaen, F., Eber, J.M. and Heath, D., Coherent measures of risk. *Math. Finance*, 1999, **9**, 203–228.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. and Shephard, N., Regular and modified kernel-based estimators of integrated variance: The case with independent noise. CAF, Centre for Analytical Finance, 2004.

Chang, C.L., Jiménez-Martín, J.Á., McAleer, M. and Pérez-Amaral, T., Risk management of risk under the Basel Accord: Forecasting value-at-risk of VIX futures. *Manage. Finance*, 2011, **37**, 1088–1106.

Chen, Q. and Gerlach, R., The two-sided Weibull distribution and forecasting financial tail risk. *Int. J. Forecasting*, 2013, **29**, 527–540.

Chen, C.W.S., Gerlach, R. and Lin, E.M.H., Volatility forecast using threshold heteroskedastic models of the intra-day range. *Comput. Stat. Data Anal.*, 2008, **52**, 2990–3010. Statistical & Computational Methods in Finance.

Chen, Q., Gerlach, R. and Lu, Z., Bayesian value-at-risk and expected shortfall forecasting via the asymmetric Laplace distribution. *Comput. Stat. Data Anal.*, 2012, **56**, 3498–3516.

Chen, C.W.S. and So, M.K.P., On a threshold heteroscedastic model. *Int. J. Forecasting*, 2006, **22**, 73–89.

Christoffersen, P., Evaluating interval forecasts. *Int. Econ. Rev.*, 1998, **39**, 841–862.

Christensen, K. and Podolskij, M., Realized range-based estimation of integrated variance. *J. Econom.*, 2007, **141**(2), 323–349.

Costanzino, N. and Curran, M., Backtesting general spectral risk measures with application to expected shortfall. *J. Risk Model Validat.*, March 2015, 21–31.

Du, Z. and Escanciano, J.C., Backtesting expected shortfall: Accounting for tail risk. Social Science Research Network Paper 2548544, 2015. Available online at: http://dx.doi.org/10.2139/ssrn.2548544.

Efron, B. and Tibshirani R.J., *An Introduction to the Bootstrap*, 1993 (Chapman and Hall: New York.

Engle, R.F. and Manganelli, S., CAViaR: Conditional autoregressive value at risk by regression quantiles. *J. Bus. Econ. Stat.*, 2004, **22**, 367–381.

Feller, W., The asymptotic distribution of the range of sums of random variables. *Ann. Math. Stat.*, 1951, **22**, 427–32.

Fissler, T. and Ziegel, J.F., Higher order eligibility and Osband's principle. *Ann. Stat.*, 2016, in press.

Fissler, T., Ziegel, J.F. and Gneiting, T., Expected shortfall is jointly elicitable with value at risk – Implications for backtesting, 2015. arXiv.org. Available online at: http://arxiv.org/abs/1507.00244.

Gaglianone, W., Lima, L., Linton, O. and Smith, D., Evaluating value-at-risk models via quantile regression. *J. Bus. Econ. Stat.*, 2011, **29**(1), 150–160.

Garman, M.B. and Klass, M.J., On the estimation of price volatility from historical data. *J. Bus.*, 1980, **53**, 67–78.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B., *Bayesian Data Analysis*, 2nd ed., 2005 (Chapman & Hall: Boca Raton, FL).

Gelman, A., Roberts, G.O. and Gilks, W.R., Efficient Metropolis jumping rules. In *Bayesian Statistics*, edited by J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, pp. 599-608, 1996 (Clarendon: Oxford).

Gerlach, R. and Chen, C.W.S., Bayesian expected shortfall forecasting incorporating the intra-day range. *J. Financ. Econom.*, 2016, **14**(1), 128–158.

Gerlach, R. and Wang, C., Forecasting risk via realized GARCH, incorporating the realized range. *Quant. Finance*, 2016, **16**(4), 501–511.

Gneiting, T., Making and evaluating point forecasts. *J. Am. Stat. Assoc.*, 2011, **106**, 746–762.

Hastings, W.K., Monte-Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 1970, **57**, 97–109.

Hansen, P.R., Huang, Z. and Shek, H.H., Realized GARCH: A joint model for returns and realized measures of volatility. *J. Appl. Econ.*, 2011, **27**(6), 877-906.

Hoogerheide, L.F., Kaashoek, J.F. and van Dijk, H.K., On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: An application of flexible sampling methods using neural networks. *J. Econom.*, 2007, **5**, 1–11.

Kerkhof, F.L.J. and Melenberg, B., Backtesting for risk-based regulatory capital. *J. Bank. Finance*, 2004, **28**, 1845–1865.

Kupiec, P.H., Techniques for verifying the accuracy of risk measurement models. *J. Deriv.*, 1995, **3**, 73–84.

Malevergne, Y. and Sornette, D., VaR-efficient portfolios for a class of super and sub-exponentially decaying assets return distributions. *Quant. Finance*, 2004, **4**, 17–36.

Martens, M. and van Dijk, D., Measuring volatility with the realized range. *J. Econom.*, 2007, **138**(1), 181–207.

McAleer, M., Jiménez-Martín, J.Á. and Pérez-Amaral, T., GFC-robust risk management strategies under the Basel Accord. *Int. Rev. Econ. Finance*, 2013, **27**, 97–111.

McNeil, A.J. and Frey, R., Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *J. Empir. Finance*, 2000, **7**, 271–300.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E., Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 1953, **21**, 1087–1092.

Molnár, P., Properties of range-based volatility estimators. *Int. Rev. Financ. Anal.*, 2012, **23**, 20–29.

Newey, W.K. and Powell, J.L., Asymmetric least squares estimation and testing. *Econometrica*, 1987, **55**, 819–847.

Parkinson, M., The extreme value method for estimating the variance of the rate of return. *J. Bus.*, 1980, **53**, 61–65.

Rogers, L.C.G. and Satchell, S.E., Estimating variance from high, low and closing prices. *Ann. Appl. Probab.*, 1991, **1**, 504–512.

So, M.K.P. and Wong, C.M., Estimation of multiple period expected shortfall and median shortfall for risk management. *Quant. Finance*, 2012, **12**(5), 739–754.

Taylor, J., Estimating value at risk and expected shortfall using expectiles. *J. Financ. Econom.*, 2008, **6**, 231–252.

Wong, W.K., Backtesting trading risk of commercial banks using expected shortfall. *J. Bank. Finance*, 2008, **32**, 1404–1415.

Yang, D. and Zhang, Q., Drift-independent volatility estimation based on high, low, open, and close prices. *J. Bus.*, 2000, **73**, 477–491.

Zhang, L., Mykland, P.A. and Aït-Sahalia, Y., A tale of two time scales. *J. Am. Stat. Assoc.*, 2005, **100**, 1394–1411.