Abstract

Bayesian semi-parametric estimation has proven effective for quantile estimation in general and specifically in financial Value at Risk forecasting. Expected shortfall is a competing tail risk measure, involving a conditional expectation beyond a quantile, that has recently been semi-parametrically estimated via asymmetric least squares and so-called expectiles. An asymmetric Gaussian density is proposed allowing a likelihood to be developed that leads to Bayesian semi-parametric estimation and forecasts of expectiles and expected shortfall. Further, the conditional autoregressive expectile class of model is generalised to two fully nonlinear families. Adaptive Markov chain Monte Carlo sampling schemes are employed for estimation in these families. The proposed models are clearly favoured in an empirical study forecasting eleven financial return series: clear evidence of more accurate expected shortfall forecasting, compared to a range of competing methods is found. Further, the most favoured models are those estimated by Bayesian methods.
Bayesian Semi-parametric Expected Shortfall Forecasting in Financial Markets

Richard H. Gerlach

Discipline of Business Analytics, University of Sydney, Australia.

(richard.gerlach@sydney.edu.au)

Cathy W. S. Chen*

Department of Statistics, Feng Chia University, Taiwan. (chenws@fcu.edu.tw)

Liou-Yan Lin

Department of Statistics, Feng Chia University, Taiwan. (ycchan@fcu.edu.tw)

*Corresponding author is Cathy W. S. Chen. Email: chenws@fcu.edu.tw.
Abstract

Bayesian semi-parametric estimation has proven effective for quantile estimation in general and specifically in financial Value at Risk forecasting. Expected shortfall is a competing tail risk measure, involving a conditional expectation beyond a quantile, that has recently been semi-parametrically estimated via asymmetric least squares and so-called expectiles. An asymmetric Gaussian density is proposed allowing a likelihood to be developed that leads to Bayesian semi-parametric estimation and forecasts of expectiles and expected shortfall. Further, the conditional autoregressive expectile class of model is generalised to two fully nonlinear families. Adaptive Markov chain Monte Carlo sampling schemes are employed for estimation in these families. The proposed models are clearly favoured in an empirical study forecasting eleven financial return series: clear evidence of more accurate expected shortfall forecasting, compared to a range of competing methods is found. Further, the most favoured models are those estimated by Bayesian methods.

Keywords: CARE model; Nonlinear; Asymmetric Gaussian distribution; Expected shortfall; semi-parametric.
1 INTRODUCTION

Value-at-Risk (VaR) is the main tool for risk measurement and subsequent capital allocation for financial institutions worldwide. However, VaR has been criticised, among other reasons, since it is not coherent, i.e. sub-additive (see Artzner, Delbaen, Eber and Heath, 1997; 1999) and also because it does not give a measure of the expected loss in the event of an extreme return. Artzner et al (1999) proposed expected shortfall (ES), also known as conditional VaR, as the expected return given that returns are more extreme than a given quantile. This measure is coherent and measures average loss for extreme events and has become widely used as a tail risk measure.

Since VaR concerns a quantile of the return distribution, semi-parametric quantile regression methods have been developed to estimate it; e.g. the CAViaR class of models proposed by Engle and Manganelli (2004). However, most ES estimation has been achieved via parametric models, with well-known conditional expectations, such as those employing Gaussian or Student-t distributions. Recently, Taylor (2008) proposed a novel semi-parametric method of estimation for ES. This method is based on the theory of expectiles, those quantities found by asymmetric least squares (ALS), and a connection between expectiles and ES found by Newey and West (1976). Taylor (2008) proposed the conditional auto-regressive expectile (CARE) class of model, estimated by ALS, then employed these to generate forecasts of ES. We extend the work of Taylor (2008) in two main directions in this paper.

Gerlach, Chen and Chan (2011) used the well-known connection between the asymmetric Laplace (AL) density and the criterion function minimised in quantile regression estimation (see Koenker and Basset, 1978) to develop a Bayesian MCMC estimator for the class of CAViaR models. Further, they extended the CAViaR family to be fully nonlinear, via the threshold (T-)CAViaR specification. In the same spirit, this paper has two main contributions. First, we extend the CARE class of models in two fully nonlinear directions, extending the symmetric absolute value CARE into a fully threshold (T-)CARE model; and extending the implied GARCH (IG-)CARE model into a threshold (T-)IG-CARE model.
Further, we propose a specific asymmetric Gaussian (AG) distribution, based on the (scaled) negative of the exponential of the ALS function employed in expectile estimation. This formulation allows a pseudo-likelihood function to be developed, whose maximum coincides with the minimum of the ALS function. In a manner analogous to that of Yu and (2001) for general Bayesian quantile regression, and also to Gerlach et al (2011) for Bayesian CAViaR estimation, this AG likelihood is then employed in a Bayesian estimation scheme.

Markov chain Monte Carlo methods, based on the adaptive mixture of Gaussian proposal method in Chen, Gerlach and Lu (2012), are employed to estimate the new and existing CARE models employed in this paper. Gerlach et al (2011) found MCMC to be more efficient at parameter estimation and more accurate at VaR forecasting, over a range of CAViaR specifications, than that obtained when numerically minimising the quantile criterion function.

The proposed threshold CARE model families, and MCMC methods developed, are examined through application to various financial market stock indices and exchange rates in a study of ES forecasting. This study illustrates that CARE models perform at least as well, and usually more accurately at forecasting ES, compared to a range of competing methods, during the period 2008-2011.

Finally, the assessment of forecasts for ES has mainly focused on (bootstrap) t-tests that a small number of residuals have mean zero. We investigate whether tests for dynamic quantiles, such as those by Christofferson (1994) and Engle and Manganelli (2004) can be usefully employed for semi-parametric ES forecasts. This follows from Chen et al (2012) who proposed such a technique for parametric ES forecast assessment.

The rest of the paper is organised as follows: Section 2 discusses expectiles, their estimation and the general link with expected shortfall, then reviews the CARE models in Taylor (2008); Section 3 introduces the proposed nonlinear extensions of the models in Section 2; Section 4 presents the asymmetric Gaussian distribution and the MCMC methods employed for estimation and forecasting; Section 5 discusses assessment of ES forecasts; Section 6 contains the empirical study and Section 7 concludes.
This section mainly reviews the work in Taylor (2008) on expectiles, CARE modelling and a link between expectiles and expected shortfall.

As discussed in Aigner, Amemiya and Poirier (1976), the $\tau$-level expectile for a random variable (r.v.) $Y$, denoted $\mu_\tau$, is a parameter that minimises the expectation $E(\mid\tau - I(Y < \mu_\tau)\mid(Y - \mu_\tau)^2)$; where $\tau \in (0, 1)$. It is important to note that in general $\mu_\tau$ is neither the $\tau$-level quantile nor the $\tau$-level expected shortfall of $Y$; in fact it does not have a simple, intuitive definition, except as the minimum of the expectation given. However, naturally $\mu_\tau$ does occur at a quantile level of $Y$, which we denote as $\alpha_\tau$: usually $\tau < \alpha_\tau$. By definition then, $\mu_\tau$ is also the minimum of $E(\left[\alpha_\tau - I(Y < \mu_\tau)\right](Y - \mu_\tau)]$. Also, setting $\tau = 0.5$ means that $\mu_{0.5}$ minimises $\text{Var}(Y)$, and so $E(Y) = \mu_{0.5}$.

Based on a sample $y_1, \ldots, y_n$ on $Y$, the $\tau$-level expectile can be estimated by minimising the asymmetric least squares (ALS) function:

$$\sum_{t=1}^{n} |\tau - I(y_t < \mu_\tau)|(y_t - \mu_\tau)^2.$$  \hspace{1cm} (1)

As shown by Newey and Powell (1987), there is a one to one connection between expectiles and expected shortfall in general. As presented in Taylor (2008), if $E(Y) = 0$ then this relationship is:

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

where all terms are defined above.

Engle and Manganelli discussed the three CAViaR models: symmetric absolute value (SAV), asymmetric (AS) and implied GARCH (IG). Taylor took the same forms as these models, replacing the dynamic quantile terms with dynamic expectile terms and proposed the three models: CARE-SAV, CARE-AS and CARE-IG, with the following specifications.

**CARE-SAV:**

$$\mu_t(\beta) = \beta_1 + \beta_2 \mu_{t-1}(\beta) + \beta_3 |y_{t-1}|$$  \hspace{1cm} (3)
where the parameters are unrestricted on the real line;

**CARE-AS:**

$$\mu_t(\beta) = \beta_1 + \beta_2 \mu_{t-1}(\beta) + (\beta_3 I(y_{t-1}>0) + \beta_4 I(y_{t-1}<0))|y_{t-1}|,$$

(4)

where the conditional expectile responds asymmetrically to positive and negative returns; and

**CARE-IG:**

$$\mu_t(\beta) = -\sqrt{\beta_1 + \beta_2 \mu_{t-1}(\beta)^2 + \beta_3 y_{t-1}^2}$$

(5)

where all parameters must be greater than 0 for positivity under the square root.

For these CARE models, Taylor (2008) estimated their unknown parameters $\beta$ by ALS, as in (1), employing the relevant conditional expectile model for $\mu_t(\beta)$. Then, equation (2) is employed to produce forecasts of ES, under an estimated $\tau$. We extend these CARE models in Section 3, to capture more flexible asymmetric and nonlinear responses, via more general threshold nonlinear forms.

### 3 PROPOSED NONLINEAR DYNAMIC EXPECTILE FAMILY

Some popular models for explaining asymmetry in the mean or volatility are the threshold autoregressive (TAR) model of Tong (1978, 1983), the threshold ARCH of Li and Li (1996) and the threshold GARCH model of Zakoian (1994); while Gerlach et al (2011) introduced the nonlinear T-CAViaR model. Motivated by these, we introduce the following threshold CARE model:

**Threshold CARE (T-CARE):**

$$\mu_t(\beta) = \begin{cases} 
\beta_1 + \beta_2 \mu_{t-1}(\beta) + \beta_3 |y_{t-1}|, & z_{t-1} \leq r \\
\beta_4 + \beta_5 \mu_{t-1}(\beta) + \beta_6 |y_{t-1}|, & z_{t-1} > r 
\end{cases}$$

(6)

Here $z$ is an observed threshold variable that could be exogenous, or self-exciting i.e. $z_t = y_t$ and $r$ is the threshold value, typically set as $r = 0$, or estimated, though empirically
Many estimates in the literature are not significant from zero; as such we fix \( r = 0 \), making the T-CARE a direct extension of the CARE-AS model in (4).

Here each parameter in the dynamic expectile function can be different in response to positive and negative responses. We call this the T-CARE family, since it includes the SAV \( (r = \infty) \) and AS \( (r = 0, \beta_4 = \beta_1 \) and \( \beta_5 = \beta_2) \) CARE models as special cases. This model is the semi-parametric equivalent of the standard deviation T-GARCH model with mean \( \mu = 0 \):

\[
\sigma_t = \begin{cases} 
\delta_1 + \delta_2 \sigma_{t-1} + \delta_3 |a_{t-1}|, & z_{t-1} \leq 0 \\
\delta_4 + \delta_5 \sigma_{t-1} + \delta_6 |a_{t-1}|, & z_{t-1} > 0 
\end{cases}
\]

(7)

We also introduce an implied GARCH threshold CARE specification, denoted T-CARE-IG model and specified as:

**Threshold Indirect-GARCH (T-CARE-IG):**

\[
\mu_t(\beta) = \begin{cases} 
[\beta_1 + \beta_2 \mu_{t-1}(\beta) + \beta_3 y_{t-1}^2]^{1/2}, & z_{t-1} \leq r \\
[\beta_4 + \beta_5 \mu_{t-1}(\beta) + \beta_6 y_{t-1}^2]^{1/2}, & z_{t-1} > r 
\end{cases}
\]

(8)

Here all parameters must be greater than 0 for positivity under the square root. Here again each parameter in the dynamic expectile function can be different in response to positive and negative responses. This model is the semi-parametric equivalent of the squared return T-GARCH model with mean \( \mu = 0 \):

\[
\sigma_t^2 = \begin{cases} 
\delta_1 + \delta_2 \sigma_{t-1}^2 + \delta_3 y_{t-1}^2, & z_{t-1} \leq 0 \\
\delta_4 + \delta_5 \sigma_{t-1}^2 + \delta_6 y_{t-1}^2, & z_{t-1} > 0 
\end{cases}
\]

(9)

The CARE-IG \( (r = \infty) \) and T-CARE-IG models form a separate branch of the threshold CARE family of models. All these CARE models are non-parametric in their error specifications, but simply extend the existing forms for the dynamic function \( \mu(\cdot) \) to be fully threshold nonlinear: i.e. they are semi-parametric nonlinear models.

Dynamic models typically have constraints or restrictions on the parameters for stationarity (or positivity of dynamic variances). However, such restrictions are difficult to locate or derive for CARE models, excepting IG models with positivity, and so we choose not to set any in this paper, as in Engle and Manganelli (2004), for the T-CARE family.
Bayesian methods generally require the specification of a likelihood function and a prior distribution. The likelihood function requires a distributional assumption to be made on the data.

### 4.1 The Asymmetric Gaussian connection

Yu and Moyeed (2001) and Tsionas (2003) illustrate the link between the solution to the quantile regression estimation problem and the likelihood for the AL distribution, allowing maximum likelihood and subsequently Bayesian estimation in quantile models. This is made possible since the AL probability density function (pdf) is the negative of the natural exponent of the quantile criterion function to be minimised. Thus any solution that minimises the criterion function is also the maximum likelihood estimator when assuming the data are from an AL distribution.

We take an analogous approach here to develop a likelihood and then a Bayesian estimator for expectile models. We propose that the kernel of a pdf for an asymmetric Gaussian (AG) rv could be the negative of the exponential of the ALS function in (1). To ensure this is a pdf it must be scaled to integrate to 1 over the real line. Doing this, we define the pdf for an AG($\tau, \mu$) as:

$$p(Y|\tau, \mu) = 2 \left[ \sqrt{\frac{\pi}{|\tau - 1|}} + \sqrt{\frac{\pi}{\tau}} \right]^{-1} \exp \left[ - (y_t - \mu)^2 / \tau I(y_t < \mu) \right],$$  

where $\mu$ is a parameter that represents the mode of the AG pdf. Figure 2 shows some examples of AG densities, clearly showing the $\mu$ is the mode of the density, while as $\tau$ gets smaller the density becomes more positively skewed; and that $\tau = 0.5$ gives the symmetric standard Gaussian.

Now, if it is assumed that the general model is given by:

$$y_t = \mu_t(\beta) + \epsilon_t$$  

$$\epsilon_t \sim AG(\tau, 0)$$
where $\epsilon_t$ is an iid process, then the corresponding likelihood function, based on a sample $y_1, \ldots, y_n$ from (11) is:

$$L_\tau(\beta; y) \propto \exp \left[ - \sum_{t=1}^{n} (y_t - \mu_t(\beta))^2 \mid \tau - I_{(-\infty, 0)}(y_t - \mu_t(\beta)) \right].$$

(13)

Since (1) is contained in the exponent of the likelihood, the mle for $\beta$ is equivalent to the expectile estimator obtained by minimising (1).

It is important to emphasize that, though we treat (13) as a likelihood function in estimation, the assumption that $y_t$ follows an $AG(\tau, \mu_t(\beta))$ distribution is not used to parametrically (indirectly) estimate expectiles or ES, nor is it used to make inferences on $\beta$. In practice, the parameter $\tau$ is fixed and/or known during estimation of $\beta$ and only that single $\tau$-level expectile of the distribution of $y_t$ is (directly) estimated here, whether employing (13) or (1). Here, (13) is employed because it leads to (ml) estimation that is mathematically equivalent to minimising (1). Use of (13), then allows a Bayesian approach, opening up powerful computational methods, such as adaptive MCMC algorithms, that employ numerical integration, which can be made arbitrarily accurate, instead of numerical optimisation. Gerlach et al (2011) found a similar approach to Bayesian quantile estimation produced more efficient parameter estimation and more accurate forecasts of
VaR (dynamic quantiles). In Section 5, we investigate the accuracy of both Bayesian forecasts of ES, produced under Bayesian estimation of $\beta$, and classical forecasts of ES, produced by estimating $\beta$ via direct minimisation of (1).

The likelihood for the T-CARE and T-CARE-IG family of models is completely specified by (13), combined with formulas for $\mu_\beta$ from either (6) or (8). We now specify the prior distribution required for a Bayesian approach.

### 4.2 Prior and Posterior densities

We choose the prior to be uninformative over the possible region for the regression-type parameters $\beta$. The joint prior is thus:

$$
\pi(\beta) \propto I(A),
$$

which is equivalent to a flat prior on $\beta$ over the region $A$. For the T-CARE model family, $A$ is equivalent to the real line for each parameter, in six dimensions. For the T-CARE-IG family, positivity is required for all parameters, so $A$ is effectively the positive real line, again in six dimensions.

Using (13) and either (6) or (8), plus the flat prior above, the joint posterior density for $\beta | \mathbf{y}$ is simply the likelihood evaluated only over the region defined by $A$. This posterior is not in the form of a known distribution in $\beta$. We thus turn to computational methods for estimation.

### 4.3 Adaptive MCMC sampling using Metropolis methods

Sampling from $p(\beta | \mathbf{y})$ directly is not possible given the non-standard form, so instead a dependent (Markov chain) Monte Carlo (MCMC) sample is obtained from $\beta | \mathbf{y}$ via adaptive versions of the Metropolis and Metropolis-Hastings (MH) (Metropolis et al., 1953; Hastings, 1970) algorithms. We employ a similar sampling algorithm to that in Chen and So (2006), employing a random walk Metropolis (RW-M) algorithm during the burn-in period and an independent kernel (IK-) MH algorithm during the sampling period.
We modify the sampling scheme of Chen and So (2006), who used Gaussian proposal densities that can get 'stuck' in local modes and be slow to move away from them, causing slow mixing and slow convergence. We employ the mixture of Gaussian proposal procedure of Chen et al (2012), with two mixture elements. Some details:

**Step 1** Burn-in RW-M with mixture proposal: $\beta|y, \tau$

$$\beta^p = \beta^{(j-1)} + \varepsilon; \quad \varepsilon \sim \rho N(0, c_1\Omega) + (1 - \rho) N(0, c_1\Omega)$$

where we set $\rho = 0.95$, $\omega = 100$ and $c_1$ is user-defined. The proposed $\theta^p$ is then accepted with the usual Metropolis probability, or rejected s.t. $\beta^p = \beta^{(j-1)}$. Step 1 is iterated, at least until convergence: these iterates make up the burn-in MCMC sample.

The MCMC sampling period employs an IKMH algorithm. First, the sample mean and sample covariance matrix of the burn-in period iterates for $\theta$ are calculated and denoted $\mu$ and $\Sigma$.

**Step 2** Sampling period Re-start the MCMC; use an IK-MH with mixture-Gaussian proposal: $\beta|y, \tau$

$$\beta^p = \gamma + \Sigma^{0.5}\varepsilon; \quad \varepsilon \sim \rho N(0, I) + (1 - \rho) N(0, I)$$

where $\Sigma^{0.5}$ is the lower triangular Cholesky decomposition, i.e. $\Sigma = (\Sigma^{0.5})'\Sigma^{0.5}$. Then accept $\theta^p$ with the usual IKMH probability.

A description follows. The scale matrix for both elements in the mixture in the burn-in period, $c_1\Omega$, which can be chosen as diagonal with positive values, is subsequently tuned to achieve optimal acceptance rates of between 15% and 50%, based on recommendations in Gelman, Roberts, and Gilks (1996) and Chen and So (2006). After the burn-in period, the sample mean vector and sample variance-covariance matrix are formed using the last $M - M_0$ iterates of $\beta$ (here $M_0$ is chosen to minimise the effect of the initial iterates of the sampling scheme; $M/5$ is usually a suitable choice). These terms are subsequently employed in the sampling period (iterations $M + 1$ to $N$) as the mean and scale matrix for another mixture of two Gaussian proposals, in an IK-MH algorithm.

This mixture of Gaussian method is a special simplified case of the more general and flexible ”AdMit” mixture of Student-t proposal procedure proposed by Hoogerheide,
Kaashoek and van Dijk (2007). We tried the AdMit method for the CARE models in this paper, but found highly similar results with those reported in Section 5.

The adaptive proposal updating procedure should speed mixing in the posterior distribution, over that for the simple RW-M method, as long as the burn-in period has ‘covered’ the posterior distribution. The mixture of Gaussian proposals employed here, deliberately chosen to have tails that are fatter than a t-distribution with low (e.g. 3-4) degrees of freedom, will greatly assist in achieving this coverage and mixing, for both the burn-in and sampling periods, by lowering the probability of getting stuck in local modes for long periods and allowing occasional ‘large’ jumps around the posterior space. We extensively examined trace plots and autocorrelation function (ACF) plots, from multiple runs of the MCMC sampler, from differing starting points, for each element of $\beta$, so as to confirm convergence and to infer adequate coverage. We also employed Gelman’s ‘R’ statistic (see pg 296, Gelman, Carlin, Stern and Rubin, 2005), which assesses speed of mixing and efficiency of convergence. The values obtained for the real data in Section 5 are typically very close to 1, and almost always below 1.05, implying fast mixing and good convergence properties, for all parameters. Observed MH acceptance rates in the sampling period for this scheme are usually from 10%-40% for all CARE models we consider.

5 ASSESSING ES FORECASTS

Once the MCMC samples are obtained for $\beta$ for a particular CARE model and data set, one-step ahead forecasts can be obtained via the dynamic expectile CARE equation, and then equation (2) is used to form a forecast of conditional ES one-step ahead.

While various common tests can be applied to directly assess VaR quantile forecasts: e.g. the unconditional coverage (UC) and conditional coverage (CC) tests of Kupiec (1995) and Christoffersen (1998) respectively, as well as the dynamic quantile (DQ) test of Engle and Manganelli (2004), the assessment of ES forecasts is still an issue under development. In particular the ES itself lies at different quantile levels for different data distributions. The main method applied for ES is based on the fact that it is a conditional expectation beyond a VaR quantile; an aspect which can be tested directly. The test examines the
residuals, observations minus predicted ES level, for data beyond the dynamic quantile VaR predictions, and tests whether residuals these could have mean 0, or not. Since the ES predictions are usually not independent over time, they are often scaled by a measure of predicted volatility, e.g. see McNeil and Frey (2000), or even scaled by the predicted VaR levels, as in Taylor (2008). We take the latter approach, since CARE models do not provide a volatility estimate, and apply the standard non-parametric bootstrap test, as in Taylor (2008), to the ES residual series divided by the VaR series. This test can have small power, however, due to the mainly small numbers of returns that are more extreme than the VaR predictions.

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. Hence, it has only been applied to fully parametric ES models, where the exact level can be calculated, see e.g. Chen et al (2012). In that case, the UC, CC and DQ tests can be applied using the quantile level of the ES, treating it as a VaR prediction at that specific quantile level. Table 1 shows the quantile levels for ES for various commonly applied distributions. These quantile levels, denoted $\delta_\alpha$, seem somewhat varied across different distributions or different degrees of freedom for a Student-t or skewed-t. However: (i) distributions with fatter tails have lower ES quantiles; (ii) the Gaussian’s tails are well-known as too thin, thus will use use lower quantile levels than Gaussian ones; (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 (0.0034, 0.0037) for a 1% ES and between (0.0175, 0.0187) for the 5% ES; (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. these ranges are very, very small for practical forecast data sizes and the tests applied to them.

To expand on the last point, if we have 1000 days in our forecast sample, as in Section 6, we can expect somewhere between 3.4 and 3.7 of these returns to be beyond our 1% ES predictions and expect between 17.5 and 18.7 returns beyond our 5% ES predictions. Thus we expect about 3-4 and 17-19 ES violating returns respectively, across a large range of conditional distributions. As such, it may be sensible for semi-parametric models to also assume that the 1% and 5% ES values will occur close enough to the 0.35% and 1.8% quantile levels, respectively, to use those quantile levels to assess the semi-parametric forecasts of ES using UC, CC and DQ tests.

The non-test criterion we use to compare ES forecasts is 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:

$$ESRate = \frac{\sum_{t=n+1}^{n+m} I(y_t < ES_t)}{m},$$

where $n$ is the estimation sample size and $m$ is the forecast sample size. As above, a series of ES forecasts should have ESRate close to the nominal level $\delta$. As also standard for VaR, the ratio ESRate/$\delta$, called the ES ratio, is employed to compare competing ES forecasts: models with ES ratio $\approx 1$ are most desirable. When ES ratio $< 1$, risk and loss forecasts are conservative (higher than actual), while alternatively, when ES ration $> 1$, risk estimates are lower than actual and financial institutions may not allocate sufficient capital to cover likely future losses. Here solvency outweighs profitability and for models where ES ratios are equidistant from 1, lower or conservative rates are preferred; e.g. ESRate/$\delta = 0.95$ is preferred to ESRate/$\delta = 1.05$, as e.g. in Wong and So (2003) and Gerlach et al (2011).

Formal testing is achieved by several well-known tests. Using either the exact value of $\delta$ for parametric models, or the approximate $\delta$ of 0.35% and 1.8% quantile levels, for 1% and 5% ES respectively, for non and semi-parametric methods, we employ the standard UC, independence, CC and DQ tests. The DQ test is well-known to be more
powerful than the CC test, see Berkowitz, Christoffersen and Pelletier (2009). These tests are now standard; we refer readers to the original papers for details. We use the extended independence and CC tests as proposed by Chen et al (2012), allowing multiple lag testing, instead of only a single lag as in the Christofferson (1998) tests.

We also apply the test of Gaglionone et al (2011), which fits a simple quantile regression model of the forecast period returns against the corresponding forecast VaR, at level $\alpha$, and testing that the parameters are 0, 1 for intercept and slope respectively, as they should be if the VaR series is accurate. We will apply this test, denoted by VQ, instead using the forecast ES as the explanatory and fitting the quantile regression at the appropriate level $\delta_{\alpha}$ for each model.

6 EMPIRICAL RESULTS

Eight daily international stock market indices are analysed: the S&P 500 (US); FTSE 100 (UK); CAC 40 (France); Dax 30 (Germany); AORD All ordinaries index (Australia); Nikkei 225 Index (Japan); TSEC weighted index (Taiwan) and the HANG SENG Index (Hong Kong). Further, three exchange rate series were considered: USDAUD, USDGBP and, USDJPY. Daily price index and exchange rate data from January 1, 2001 to September 5, 2011 were obtained from Yahoo Finance. The percentage log return series were 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 $t$.

The full data period is divided into a learning or estimation sample: January 1, 2001 to December 31, 2007; and a forecast sample: the approximately 1000 trading days from January 1, 2008 to September 5, 2011. Small differences in forecast sample sizes and end-dates occurred across markets due to different market-specific non-trading days. All eleven series display the standard properties of daily asset returns: they are heavy-tailed and mostly negatively skewed.

Parameter estimates from the CARE models are not shown to save space. However, for each dataset, the MCMC burn-in sample size is 8,000 iterations, followed by a sampling period of 2,000 iterations. To assess mixing and convergence, the MCMC method is
run from five different, randomly generated starting positions, for each market, at $\alpha = 0.01, 0.05$. The starting values for each of the parameters, i.e. $(\beta_1, \ldots, \beta_6)$ were chosen to be widely varying and to lie on either side of the estimated posterior mean. Convergence to the same posterior distribution is clear in all five runs for each parameter, in each case well before the end of the burn-in sample, in each market. Gelman’s R statistics (see Gelman et al., 2005, page 296) over these five runs, for the UK market data, were between 1.002 and 1.040 over the six parameters: a typical result; all highlighting fast mixing and clear and efficient convergence for the proposed sampling scheme. Further, observed MH acceptance rates in these MCMC runs were between 15% and 60%. These figures are quite representative of the R values and observed MH acceptance rates during the sampling period across all markets and CARE models.

6.1 Forecasting ES study

ES is forecast 1 day ahead for each day in the forecast sample of $\approx 1000$ returns, in each return series, using a range of competing models, including non-parametric, semi-parametric and fully parametric 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. We take this approach for two commonly used sample sizes: the last 100 days (HS100) and a longer-term last 250 days (HS250, approximately one trading year). Four CARE models are considered, the full self-exciting T-CARE (TC) in (??), as well as the SAV, AS and IG CARE models given in (3), (4) and (5) respectively. These models are estimated in turn by the Bayesian MCMC method (e.g. IG-B) proposed here, and then by classical ALS minimisation, using the optimization toolbox in Matlab. A range of popular GARCH specifications, including with Gaussian errors (G-n), Student-t errors (G-t) and skewed-t errors (G-st) as in Hansen (1994) is also considered. These are estimated by ml, using the Econometrics toolbox in Matlab, while the GARCH-skt uses programs adapted from the website of Kevin Sheppard, xxx. Finally an extreme value is considered, using the peaks-over-threshold (POT) method on returns standardised by volatility estimated by the RiskMetrics method; see Galli and Kellezi (2006) for details.
Figure 2: Some 1% ES forecasts and the UK FTSE100 forecast sample returns.

For each day $y_{n+t}$ in the forecast sample, parameters are estimated for each semi-parametric and parametric model, employing the fixed window size $n$ of data $(y_t, \ldots, y_{n+t-1})$ as observations. Forecasts are then calculated for the next day’s $\alpha$-level ES. Under ml or ALS estimation, the parameter estimates are simply substituted into each forecast equation for ES under each model, to generate ES forecasts. Under Bayesian estimation for the CARE models, each set of parameter iterates is substituted into the CARE equation to generate an expectile forecast that is transformed into an ES forecast. Thus, 25,000 (post burn-in) MCMC ES forecast iterates are generated each day for each CARE model: these are simply averaged to give the final ES forecast. Each MCMC run takes about 10 seconds, while estimating $\tau$ by grid search takes a few seconds, on a standard modern desktop PC; thus estimating 1000 days of forecasts takes a matter of a few hours, for each CARE model at each level.

Figure 2 shows the returns in the forecast period for the UK market, as well as the 1% ES forecasts for five different methods: G-n, G-t, SAV-B, HS250 and G-AL. The forecasts for the GARCH model with Gaussian errors (G-n) are usually the least extreme for each day, as is well-known and expected, while forecasts for the G-t and CARE-SAV, under Bayesian estimation (SAV-B), seem quite close on most days. On the contrary, forecasts using the GARCH with AL distributed errors are usually, and clearly, the most extreme each day. The AL is estimating much fatter tails than the Student-t distribution here. The HS250 method is clearly affected adversely by extreme returns, causing a flat
Figure 3: Some 1% ES forecasts and the USJP exchange rate forecast sample returns.

forecast for up to 250 days, while also responding most slowly to extreme returns or increased volatility periods. Figure 3 shows the forecast period returns for the USJP exchange rate, and 1% ES forecasts for the G-n, G-st, POT, TC and TC-B models. Here the G-n and POT are mostly the least extreme forecasts, with the occasional brief period where the ALS estimated TCARE model forecasts are least extreme. In particular, there are two periods in the last 250 days where the ES forecasts from the TC=ALS model get inexplicably and sub-optimally close to 0: this causes problems that will become clear below. The TC-B forecasts during the same periods do not display the same poor behavior at all, while being quite close to the TC-ALS forecasts for most other periods. The G-st method forecasts are more extreme than the G-n, as expected, and are similar in level, though smoother, than those for the TC-B model.

The general patterns for and among the forecasts for the different methods are fairly consistent across the eleven return series, for both 1% and 5% ES forecasting. To summarise, usually the G-n ES forecasts are the least extreme in each market on most days; the CARE model’s forecasts, as a group, are similar to the G-t and G-st on most days, while the HS100 and HS250 ES forecasts are consistently and completely different to the other methods. Further, the behavior of the POT method ES forecasts is quite inconsistent compared to the other methods across. For some return series it is the most extreme of all ES forecasts, e.g. for the UK (not shown in Figure 2 since it distorts the picture, while for some other series, like the US/JP exchange rate, it is the least extreme of the ES
forecasts; over the eleven series the forecasts from the POT method are the most variable, in comparative level, with the other methods. ES forecasts from the CARE models are mostly very similar to each other, regardless of ALS or Bayesian estimated. However, in each series there are small periods where ALS and Bayesian estimated CARE models diverge in their forecasts, this is more apparent for 1% ES forecasting, as shown for the USJP returns in Figure 3; in those periods the Bayesian estimated models appear to give more sensible ES forecasts.

Figure 4 shows the average ES residual in each of the eleven return series, for each of the models/methods, from the 1% ES forecasts. These averages should (theoretically) be zero for an accurate ES forecast method. Each ‘+’ is the average ES residual for a particular return series or market and model/method, while each ‘o’ shows the mean of the eleven ES residual averages for each model/method; a reference line is at zero. Clearly, all of the methods under-predict, on average, the true ES level, across the eleven series, when there is a return beyond the VaR quantile, since all averages are negative. This indicates that most methods are anti-conservative, since hey under-estimate the true loss or risk level, on average. However, most of the models have at least one market where the average ES residual is positive; excepting the G-n, HS100 and ALS estimated TC model, which are clearly and significantly biased in their ES forecasting in all eleven series. The three models with mean of average ES residuals closest to 0 are the SAV-B (Bayesian estimated), G-t and AS-B, followed closely by the ALS estimated SAV model. The variability across markets in the level of ES forecasts by the POT method is clear in this figure, while the problematic close to 0 ES forecasts by the TC-ALS model in USJP return series causes a very large negative ES residual, not shown in Figure 4 since it is off the scale of the y-axis, causing the very large negative mean ES average residual for that method. Among the CARE models, those estimated by Bayesian methods have mean ES residuals closer to 0 than those estimated by ALS, excepting the CARE-IG model.

Figure 5 shows the ESRates for the 1% ES forecasts for each model in the forecast period over the eleven series. Each ‘+’ is the ESRate for a particular return series and model/method (marked on x-axis). These rates should be close to the $\delta_{alpha}$ for each model, which should be between 0.0035 and 0.0038; reference lines are shown for these
Figure 4: Averages of the ES residuals, for 1% ES forecasts, for each of the eleven return series, for each model/method (marked on x-axis). The reference line is at 0, where accurate models are expected to have their average ES residual; 'o' indicates the mean of the eleven ES residual averages for each model.

Figure 5: ESRates, for 1% ES forecasts, across all eleven series for each model/method considered. Reference lines are at 0.0035 and 0.0038: models are expected to have true nominal ES quantile levels in this range.
values. Clearly, most of the methods have ESRates well above their nominal level(s), which is expected since the ES residuals average below zero: we know the forecasts are generally not extreme enough. In fact, for most models the ES forecasts would seem to be more appropriate as 1% VaR forecasts, which is clearly problematic: we expect the 1% ES to be well below the 1% VaR for most well-behaved distributions. The models doing the best under this metric are the G-t, G-st, SAV, SAV-B, AS-B and TC-B models. All other methods are significantly anti-conservative, having far too many ES violations in all or most return series.

We now consider formally testing the conclusions above and each model’s ES forecasts for accuracy and independence, across the eleven series. Table 2 counts the number of rejections, at the 5% significance level, for each model/method, over the eleven return series, for each of the tests considered: bootstrap t-test, UC, independence (4 lags), CC (4 lags), DQ (for lags) and VQ tests. Four lags were used for the extended independence and CC, as well as the DQ, tests as in Engle and Manganelli (2004). For 1% ES forecasting, clearly the G-n, POT, HS100, HS250, IG and IG-B are the worst performing models, being rejected in almost all the markets by three or four of the tests. The bootstrap and lag 4 independence test clearly reject the least number of models, as expected since these tests are known to have low power; in particular the bootstrap t-test is being applied to about 10 observations only, since only 10 returns are expected to be more extreme than their corresponding 1% VaR forecasts. The lag 4 CC test and DQ tests reject similar numbers of models in each market, while the VQ typically rejects slightly fewer models in each market, a surprising result given the power study in Galglionone et al (2011). Only four models are not rejected in most markets: the G-t, SAV-B, AS-B and TC-B, the latter three all CARE models estimated by Bayesian methods; These CARE models are each rejected in only four markets, by the DQ test, and in less markets by all other tests. All other models are rejected in at least five or six (i.e. most) markets by at least one of the tests.

A similar story applies to the 5% ES forecasts, in regards to the test results. However, now only the SAV-B and AS-B models are not rejected in most of the markets; the G-t is rejected in nine series by the DQ test, while the TC-B is rejected in six series by the DQ.
Table 2: Counts of model rejections for six formal ES forecast assessment tests, across the eleven markets.

<table>
<thead>
<tr>
<th></th>
<th>G-n</th>
<th>G-t</th>
<th>G-st</th>
<th>G-AL</th>
<th>HS100</th>
<th>HS250</th>
<th>POT</th>
<th>IG</th>
<th>IG-B</th>
<th>SAV</th>
<th>SAV-B</th>
<th>AS</th>
<th>AS-B</th>
<th>TC</th>
<th>TC-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 0.01$ bootstrap</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IND4</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>UC</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CC4</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DQ4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>11</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>VQ</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>G-n</th>
<th>G-t</th>
<th>G-st</th>
<th>G-AL</th>
<th>HS100</th>
<th>HS250</th>
<th>POT</th>
<th>IG</th>
<th>IG-B</th>
<th>SAV</th>
<th>SAV-B</th>
<th>AS</th>
<th>AS-B</th>
<th>TC</th>
<th>TC-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 0.05$ bootstrap</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>IND4</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>UC</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CC4</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DQ4</td>
<td>10</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>11</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>VQ</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: Boxes indicate the favoured model, bold indicates the least favoured model, in each column.
Figure 6: Averages of the ES residuals, for 5% ES forecasts, for each of the eleven return series, for each model/method (marked on x-axis). The reference line is at 0, where accurate models are expected to have their average ES residual; ‘o’ indicates the mean of the eleven ES residual averages for each model test. Figures 6 and 7 show the average ES residuals, and the ES violation rates (ESRates) in the eleven series for each model/method. Clearly, the most accurate 5% ES forecasts on average are from both SAV models, followed by the AS-B, TC-B and G-t. However, clearly no model or method is particularly close to the expected nominal ESRates of 0.0175 – 0.0185 (or 0.0196 for Gaussian) when forecasting 5% ES during this time period, excepting the GARCH model with AL errors. However, despite having the only positive mean ES residual, indicating consistently conservative ES forecasts, and having ESRates closest to nominal over the eleven series, this model is rejected in eight of these series by the DQ test, and in seven series by the independence test: clearly it does not capture the dynamics of 5% ES levels in most markets. The other models consistently forecast 5% ES to be between the 2% and 3% quantile levels, above what is expected.

In summary, for 1% ES forecasting, the CARE family of models are highly competitive at worst, and far more accurate at best, at dynamic ES forecasting, compared to a wide range of popular and well-known ES forecasting methods. The SAV-B, AS-B and TC-B models performed similarly accurately at 1% ES forecasting and were rejected in the least number of series. The G-t method performed only slightly worse, while all other methods considered were rejected in most of the series considered by at least test, and had ES violation rates at or above 1%, clearly too high for 1% ES forecasts.

21
For 5% ES forecasting, the GARCH-AL gave consistently conservative forecasts and had ESRates closest to nominal (0.0184), however was rejected in most markets regarding conditional coverage. The other markets were mostly similar in ESRates, while again the CARE models under Bayesian estimation, the SAV-B and AS-B, fared best under the battery of tests applied, being the only models not rejected in most markets.

7 CONCLUSION

This paper considers dynamic expectile and expected shortfall modelling and forecasting. Expectiles are estimated using an asymmetric least squares function. An asymmetric Gaussian distribution is developed to use as the errors in an expectile model, leading to a likelihood formulation whose maximum occurs at the minimum of the asymmetric least squares function. Bayesian MCMC methods are employed in this framework to estimate the class of CARE models. This class is extended to two fully nonlinear families. An expected shortfall forecasting study, using data during and after the global financial crisis, reveals that the T-CARE model family under Bayesian estimation compares most favourably in terms of ES residuals, violation rates and independence of violations, to the CARE-IG model, T-CARE models estimated via direct asymmetric least squares and to a range of well-known GARCH models, historical simulation and an extreme value method,
across eleven return series.

**Acknowledgements** Richard Gerlach is supported by ARC grants DP0985505 and DP1094848.

**References**


Gerlach, R, Lu, Z and Huang (2011), “,” University of Sydney working paper no.


Koenker, R and Machado (1999)


25


