The University of Sydney Business School
The University of Sydney

BUSINESS ANALYTICS WORKING PAPER SERIES

Bayesian Assessment of Dynamic Quantile Forecasts

Richard Gerlach[1], Cathy W.S. Chen[2] and Edward M.H. Lin[3]

[1] The University of Sydney Business School, Australia.
[2] Department of Statistics, Feng Chia University, Taiwan.
[3] Academica Sinica, Taiwan.

Abstract

Methods for Bayesian testing and assessment of dynamic quantile forecasts are proposed. Specifically, Bayes factor analogues of popular frequentist tests for independence of violations from, and for correct coverage of a time series of, quantile forecasts are developed. To evaluate the relevant marginal likelihoods involved, analytic integration methods are utilised when possible, otherwise multivariate adaptive quadrature methods are employed to estimate the required quantities. The usual Bayesian interval estimate for a proportion is also examined in this context. The size and power properties of the proposed methods are examined via a simulation study, illustrating favourable comparisons both overall and with their frequentist counterparts. An empirical study employs the proposed methods, in comparison with standard tests, to assess the adequacy of a range of forecasting models for Value at Risk (VaR) in several financial market data series.

Keywords Bayesian Hypothesis testing; Bayes factor; asymmetric-Laplace distribution; Value-at-Risk; quantile regression.

September 2014

BA Working Paper No: BAWP-2014-04
http://sydney.edu.au/business/business_analytics/research/working_papers

Bayesian Assessment of Dynamic Quantile Forecasts

Richard Gerlach[1], Cathy W.S. Chen[2*]and Edward M.H. Lin[3]

[1] The University of Sydney Business School, Australia.

[2] Department of Statistics, Feng Chia University, Taiwan.

[3] Academica Sinica, Taiwan.

Abstract

Methods for Bayesian testing and assessment of dynamic quantile forecasts are proposed. Specifically, Bayes factor analogues of popular frequentist tests for independence of violations from, and for correct coverage of a time series of, quantile forecasts are developed. To evaluate the relevant marginal likelihoods involved, analytic integration methods are utilised when possible, otherwise multivariate adaptive quadrature methods are employed to estimate the required quantities. The usual Bayesian interval estimate for a proportion is also examined in this context. The size and power properties of the proposed methods are examined via a simulation study, illustrating favourable comparisons both overall and with their frequentist counterparts. An empirical study employs the proposed methods, in comparison with standard tests, to assess the adequacy of a range of forecasting models for Value at Risk (VaR) in several financial market data series.

Key words:Bayesian Hypothesis testing; Bayes factor; asymmetric-Laplace distribution; Value-at-Risk; quantile regression.

1 Introduction

A wealth of recent interest in dynamic quantile modelling and forecasting creates a demand for tests and methods to assess the accuracy of quantile predictions. A prime

*Corresponding author: Cathy W.S. Chen, Email:chenws@mail.fcu.edu.tw

example from the financial markets is Value-at-Risk (VaR), which corresponds to the multiple of a quantile of the financial return distribution. Following the Basel II Capital Accord, VaR is widely used in practice for risk management and capital allocation, to protect against large negative market movements in asset prices. VaR is now the primary financial risk measure used in banking and by financial institutions. The accord further advises risk managers to regularly back-test the VaR methods being used, using at least one year of historical data to compare actual returns with VaR forecasts.

Four well-known formal back-testing methods for quantile (VaR) forecasts are the unconditional coverage (UC) test of Kupiec (1995), the conditional coverage (CC) test of Christoffersen (1998), the dynamic quantile (DQ) test of Engle and Manganelli (2004), and the VaR Quantile Regression (VQR) test of Gaglianone et al. (2011). Berkowitz, Christofferson and Pelletier (2011) developed a unified Lagrange Multiplier framework for VaR assessment, incorporating these tests (except the VQR). Gaglianone et al. (2011) highlighted that the VQR and DQ tests were over-sized in general, more-so in smaller samples, but have higher size-adjusted power than the UC and CC tests. The latter outcome highlights that the VQR and DQ tests use more information, and in a mostly more effective manner, than the binary variables which the UC and CC tests solely rely on; whilst the former result indicates a potential opportunity to develop better tests, being the goal of this paper.

Bayesian methods are widely employed for forecasting Value at Risk (VaR); see e.g. Hoogerheide and van Dijk (2010), Chen et al. (2012a, 2012b) and Gerlach and Chen (2008), etc. However, Bayesian methods for back-testing are not prevalent, or apparently even existing, in the literature. Thus, such Bayesian papers typically revert to frequentist tests to assess and compare VaR models. The major goal of this paper is to fill this gap in the literature by proposing formal back-testing methods for dynamic quantile predictions that are within a Bayesian framework. This is achieved by developing a suite of Bayesian methods that are roughly analogous to the existing frequentist tests, as well as employing some common Bayesian methods that are not commonly applied in this area. Most of the proposed methods are based on Bayes factors, which require estimation of marginal likelihoods, for which we suggest analytic methods when feasible, otherwise we employ either multivariate quadrature methods or other approximation to estimate

these quantities.

A second goal of this paper is to assess whether the proposed Bayesian testing framework, that does not rely on large sample or asymptotic approximations to null distributions, could be more effective in testing VaR, and other dynamic quantile, forecasting methods; as judged by their sampling properties. As with the UC, CC, DQ and VQR tests, the proposed Bayesian tests will not depend on the model that generated the data, nor on the method of estimation of the model parameters involved in forecasting dynamic quantiles. In contrast to the common situation of Bayesian forecasting methods being assessed via frequentist tests, the proposals in this paper will allow both Bayesian and frequentist forecasting methods to be assessed via Bayesian tests.

The article is organized as follows: Section 2 reviews the evaluation of quantile forecast accuracy via existing tests; Section 3 introduces Bayesian hypothesis testing via various test procedures; Sections 4 and 5 present and further discuss the results from a simulation; Section 6 illustrates an empirical study with a range of competing VaR methods; while concluding remarks appear in Section 7.

2 Evaluating quantile forecast accuracy

VaR is now a standard tool in risk management. It is an estimate (forecast) of the size of the minimum potential loss, over a given time horizon, with a specified probability, for a financial position. Let y_t denote the return observation at time t, then VaR (VaR_t) at level α can be defined via:

$$\Pr(y_t < -\text{VaR}_t|\mathcal{F}_{t-1}) = \alpha,$$

where $\mathcal{F}_{t-1}$ is the past information, available at time $t-1$. For a forecast sample period, the observed violation rate is the number of violations, i.e. return observations that are more extreme than their respective VaR forecast ($y_t < -\text{VaR}_t$) $= 1$), divided by the forecast sample size n.

Kupiec (1995)'s likelihood ratio (LR) UC test examines the hypothesis that the true violation rate is equal to α, as required for an accurate VaR forecasting method. The LR

test statistic is:

$$LR_{uc} = 2\left\{\log[\hat{\alpha}^{n_1}(1-\hat{\alpha})^{n-n_1}] - \log[\alpha^{n_1}(1-\alpha)^{n-n_1}]\right\},$$

where n_1 is the number of violations in n observations and $\hat{\alpha}$ is the observed sample violation rate. This assesses whether the binary violation indicator series $I(y_t < -\text{VaR}_t)$, $t = 1, \ldots, n$ could have an incidence rate equal to α, or not. Under the null, which also assumes the binary series is i.i.d. Bernoulli, LR_{uc} tends towards a $\chi^2(1)$ distribution as n gets large.

Christoffersen (1998) develops a conditional coverage (CC) joint test, incorporating the UC test, that the binary violations are independent and occur with nominal rate α; the joint LR test is:

$$LR_{cc} = -2\log[\alpha^{n_1}(1-\alpha)^{n-n_1}] + 2\log[(1-\pi_{01})^{n_{00}}\pi_{01}^{n_{01}}(1-\pi_{11})^{n_{10}}\pi_{11}^{n_{11}}]$$

where n_{ij} is the number of occurrences of $I_t = j, I_{t-1} = i$, for $i, j = 0, 1$, where $I_t = I(y_t < -\text{VaR}_t)$. Under the alternative, the violations follow a two-state Markov chain process, where $\pi_{ij} = m_{ij}/\sum_j m_{ij}$, $i, j = 0, 1$, are the observed Markov transition rates, which violates the independence assumption of the null. Under the null, LR_{cc} tends towards a χ^2_2 distribution as n gets large. The result $LR_{cc} = LR_{UC} + LR_{ind}$, where LR_{ind} is the independence test of Christoffersen (1998), follows from the definitions of these test statistics.

Engle and Manganelli (2004) develop the DQ test, another joint test for correct coverage and independence, but one that can employ more than just the binary violation series. The null is $H_0 : I(y_t < -\text{VaR}_t)$ are an i.i.d. series with rate α. A series of "hits", $H_t = I_t - \alpha$, are then calculated. Under the null it is straightforward to show that $E(H_t) = 0$ and $E(H_t W_{it}) = 0$, where W contains q relevant explanatory variables that are in the information set at time $t-1$, when the forecast VaR_t is made. The DQ test statistic examines whether all parameters in a regression of H on W equal zero, calculated as:

$$\text{DQ}(q) = \frac{\boldsymbol{H'W}\left(\boldsymbol{W'W}\right)^{-1}\boldsymbol{W'H}}{\alpha(1-\alpha)},$$

which is analogous to a regression F statistic. Under the null, $DQ(q)$ tends to a χ^2_q distribution as n gets large. As in Engle and Manganelli (2004), we employ lagged hits

and the VaR forecast itself as explanatory variable choices, i.e. $W_t^{'} = (1, H_{t-1}, \text{VaR}_t)$ (denoted as DQ1) and $W_t^{'} = (H_{t-1}, \ldots, H_{t-4}, \text{VaR}_t)$ (denoted as DQ4).

Gaglianone et al. (2011) employed direct "Mincer-Zarnowitz" quantile regression for the αth conditional quantile of y_t:

$$Q_{y_t}(\alpha|\mathcal{F}_{t-1}) = \beta_0 + \beta_1 VaR_t, \quad \text{for all } \alpha \in (0,1),$$

subsequently developing a relevant test to assess forecast accuracy for a VaR method; their test employs the actual data, not violation indicators, as well as the VaR forecast series itself, thus employing more information than the UC, CC and DQ tests. If the VaR forecasts are accurate, then $\beta_0 = 0, \beta_1 = -1$ in this quantile regression specification. Gaglianone et al. (2011) test that hypothesis, i.e. $\theta = (\beta_0, \beta_1 - 1)^{'} = \mathbf{0}$ using the statistic $\text{VQR} = \hat{\theta}^{'} \left(\hat{\Sigma}\right)^{-1} \hat{\theta}$, that asymptotically follows a χ_2^2 under the null hypothesis. We followed Gaglianone et al. (2011) and Koenker and Machado (1999)'s recommendations here, in particular in estimation of Σ and its components. See those papers for details.

These are the four most commonly applied tests to assess the accuracy of quantile forecasts. Though they employ different information sets, we will employ them as recommended by their author developers. In the next section we address the gap in the literature regarding Bayesian assessment of quantile forecasts.

3 Bayes Factors and Hypothesis Testing

In a Bayesian framework, hypothesis testing and model comparison problems can be tackled via posterior credible intervals or via marginal likelihoods that are often translated into Bayes factors (BFs). BFs are estimated via marginal likelihoods: $p(y|M_k) = \int p(y|\theta, M_k) p(\theta|M_k) d\theta$ where model M_k is generally preferred over M_j if $\text{BF} = \frac{p(y|M_k)}{p(y|M_j)} > 1$. BFs can also be employed in hypothesis testing of $\theta = \theta_0$, where the hypothesis is rejected if $\frac{p(y|\theta_0, M)}{p(y|M)} < 1$.

3.1 Bayesian testing for unconditional and conditional coverage

The null hypothesis in the UC test is $H_0 : \alpha = \alpha^*$, where α^* is the nominal quantile level. This hypothesis can be directly tested using a Bayesian credible interval for α.

Under an assumed binomial Bin(m, α) distribution for the number of violations n_1, and employing a conjugate Beta(a, b) prior, a Beta$(n_1 + a, n - n_1 + b)$ posterior distribution results for $p(\alpha|I)$. As standard, we choose both a flat Beta$(a = 1, b = 1)$ prior and the Jeffreys' Beta$(a = 0.5, b = 0.5)$ prior, and simply form the 95% posterior credible interval from the resulting Beta$(n_1 + a, n - n_1 + b)$ distribution, employing the 2.5th and 97.5th quantiles in each case. The hypothesis $\alpha = \alpha^*$ is rejected whenever α^* is outside the obtained credible interval; we label these methods "Bp11" and "Bp55" respectively. See Gelman et al. (2005, Chapter 2), Tuyl, Gerlach and Mengersen (2008) and Brown et al (2001), for more details and discussion on Bayesian inference for proportions and the frequentist properties of such.

A BF test, roughly analogous to the frequentist likelihood ratio UC test is also proposed, where:

$$\text{BFUC} \quad = \quad \frac{\alpha^{n_1}(1-\alpha)^{n-m_1}}{\int \pi^{n_1}(1-\pi)^{n-n_1}p(\pi)d\pi}$$

involves the Binomial likelihood evaluated under the null, divided by the marginal likelihood, where $\pi \sim$ Beta(a,b) and where $a = b = 1$ is chosen, as standard. Then, $H_0 : \alpha = \alpha^*$ is rejected whenever:

$$\text{BFUC} = \frac{\alpha^{n_1}(1-\alpha)^{n-n_1}}{B(n_1+1, n-n_1+1)} < 1 \ ,$$

where $B(c, d) = \frac{\Gamma(c)\Gamma(d)}{\Gamma(c+d)}$ is the standard incomplete Beta integral and Γ is the standard Gamma function.

BF analogues are also developed for the independence and CC tests of Christofferson (1998). First, the null model is $M_0 : I_t \sim$ i.i.d. Binomial(n, α) vs $M_1 : I_t|I_{t-1} \sim$ Binomial$(n, \pi_{i,j})$, $i, j = 0, 1$, where the alternative is a two-state Markov chain and $\pi_{i,j} = Pr(I_t = j, I_{t-1} = i)$. Ignoring the combinatorial terms, as in Christofferson (1998), and after integration, the BF is the ratio of marginal likelihoods:

$$\text{BFind} \quad = \quad \frac{B(n_1+1, n-n_1+1)}{B(n_{01}+1, n_{00}+1)B(n_{11}+1, n_{10}+1)}$$

where the null model M_0 is rejected whenever BFind < 1. For the CC BF method we have:

$$\text{BFCC} \quad = \quad \frac{\alpha^{n_1}(1-\alpha)^{n-n_1}}{B(n_{01}+1, n_{00}+1)B(n_{11}+1, n_{10}+1)}$$

where n_{ij} is the number of instances where $I_t = j, I_{t-1} = i$ for $i, j = 0, 1$ and $t = 2, \ldots, n$. Analogous to the relationship between the UC, independence and CC LR tests, here BFCC = BFUC $\times$ BFind. The BFUC involves the binomial likelihood evaluated under the null, divided by the marginal likelihood. The null model M_0 is rejected whenever BFCC < 1.

3.2 Bayesian DQ testing

A Bayes factor requires an assumed model and data distribution to produce a likelihood. The DQ test employs the series of "hits" $H_t = I_t - \alpha$, $t = 1, \ldots, n$ and fits a regression:

$$H_t = \beta_0 + \sum_{i=1}^{(q-1)} \beta_i W_{i,t} + \epsilon_t$$

To get a likelihood, a distribution can be assumed for ϵ_t. The simplest, but admittedly non-intuitive, choice is $\epsilon_t \sim N(0, \sigma^2)$. This leads to

$$p(H|\beta, \sigma^2) = (2\pi)^{-0.5(n-q-1)} \sigma^{-\frac{n-q-1}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=q+1}^{n} \epsilon_t^2\right) .$$

σ^2 is a nuisance parameter here, but under the standard Jeffreys' prior $p(\sigma^2) \propto \sigma^{-2}$ it can be analytically integrated out, giving:

$$\text{BFDQ} = \frac{\left[0.5 \sum_{t=q+1}^{n} H_t^2\right]^{-m/2}}{\int \ldots \int \left[0.5 \sum_{t=q+1}^{n} \epsilon_t^2\right]^{-n/2} p(\beta) d\beta}.$$

Under a proper Gaussian prior on β, e.g. $\beta|\sigma^2 \sim N(0, C\sigma^2)$ (where C is a diagonal matrix with large elements), the denominator can be integrated analytically (e.g. as in Smith and Kohn, 1996) and BFDQ calculated. Under the null all $\beta = 0$, which is rejected whenever BFDQ< 1. We employ the same regressors as in the DQ statistics, giving BFDQ1 and BFDQ4 procedures.

A more intuitive BF method, also analogous to the DQ test, is obtained via a standard logistic regression. Here we let:

$$Pr(I_t = 1|W_t) = \text{logit}\left[\beta_0 + \sum_{i=1}^{(q-1)} \beta_i W_{i,t}\right]$$

, where $\text{logit}(x) = (1 + \exp(-x))^{-1}$. The null hypothesis has $\beta_0^* = \log\left(\frac{\alpha}{1-\alpha}\right)$ and $\beta_i^* = 0; i = 1, \ldots, q-1$. The BFLDQ statistic is formulated as:

$$\text{BFLDQ} = \frac{p(I|\beta = \beta^*)}{\int p(I|\beta)p(\beta)d\beta}.$$

A proper Gaussian prior on β is employed, e.g. $\beta|\sigma^2 \sim N(0, C\sigma^2)$ (where C is a diagonal matrix with large elements) to evaluate the denominator. However, this prior is not conjugate and the integral cannot be evaluated analytically using known methods. To estimate this integral, the method in Kass and Raftery (1995) based on approximating the integrand by a second order Taylor series expansion and then analytically integrating the resulting Gaussian density function is employed. This method leads to:

$$\int \exp(-g(\beta)2)d\beta \approx (2\pi)^{0.5q} \exp(-g(\hat{\beta})2) \left| g^{''}(\beta)/2 \right|^{0.5}_{\beta=\hat{\beta}},$$

where $g(\beta) = -2\log\left(p(I|\beta)p(\beta)\right)$. The term $g^{''}(\beta)$ is the matrix of 2nd derivatives of $g(.)$, which if $X_t = (1 W_{1,t} \ldots W_{q-1,t})$, is given by:

$$\sum_{t=1}^{n} X_t X_t^{'} \times \text{logit}(X_t\beta)\left(1 - \text{logit}(X_t\beta)\right).$$

Here again the same regressors as in the DQ statistics are used, with focus on $q = 2$ (1 lag) and $q = 5$ (4 lags), giving the BFLDQ1 and BFLDQ4 statistics.

3.3 Bayesian VQR testing

Koenker and Machado (1999) note that quantile regression estimation, usually performed by minimising the quantile distance function:

$$\min_{\boldsymbol{\beta}} \sum_t u_t \left[\alpha - I(u_t < 0)\right]$$

is equivalent to a maximum likelihood (ML) estimation procedure when assuming i.i.d. skewed Laplace errors, i.e. $u \sim SL(0, \sigma, \alpha)$, so that:

$$p_\alpha(u) = \frac{\alpha(1-\alpha)}{\sigma} \exp\left[-\left(\frac{u\left[\alpha - I(u < 0)\right]}{\sigma}\right)\right].$$

The ML and usual quantile regression estimates for $\boldsymbol{\beta}$ are mathematically equivalent in this case.

The quantile regression model for the αth conditional quantile of y_t, regressed against its VaR forecast, can be written:

$$Q_{y_t}(\alpha|\mathcal{F}_{t-1}) = \beta_0 + \beta_1 VaR_t, \quad \text{for all } \alpha \in (0,1).$$

If the VaR forecasts are accurate, then the parameters should conform with $\beta_0 = 0, \beta_1 = 1$, as assessed by the VQR test of Gaglianone et al. (2011).

For the BFVQ procedure, again assuming a Jeffreys prior on σ and integrating it out gives:

$$p(\boldsymbol{u}|\boldsymbol{\beta}) = \alpha^n(1-\alpha)^n\Gamma(n)\left[\sum_{t=1}^{n} u_t(\alpha - I(u_t < 0))\right]^{-n} .$$

Thus, the BFVQ statistic is:

$$\text{BFVQ} = \frac{p(\mathbf{u}|\beta_0 = 0, \beta_1 = 1)}{\int\int p(\mathbf{u}|\beta)p(\beta)d\beta_0 d\beta_1}$$

where the null of $\beta_0 = 0, \beta_1 = 1$ is rejected whenever $BFVQ < 1$. The denominator above is a double integral over the bivariate real line. We employ a diffuse, proper Gaussian prior on β, then transform to the region $(-1,1) \times (-1,1)$ and use adaptive quadrature methods to numerically estimate this integral. This takes less than half a second on a standard laptop using Matlab software and function "dblequad".

4 Simulation study

The empirical properties of the proposed Bayesian methods are assessed via a simulation study. The same simulation setting as in Gaglianone et al. (2011) is employed. The true model is a GARCH(1,1), specified as:

$$\sigma_t^2 = 0.1 + 0.1y_{t-1}^2 + 0.85\sigma_{t-1}^2 \; ; \; y_t = \sigma_t\epsilon_t \; ; \; \epsilon_t \sim N(0,1)$$

where $\text{VaR}_{t,\alpha} = \sigma_t\Phi^{-1}(\alpha)$. To assess power an incorrect, but common, historical simulation (HS) VaR estimator is employed:

$$\text{HS250}_{t,\alpha} = \hat{Q}_\alpha(y_{t-250}, \ldots, y_{t-1})$$

using the sample percentile of the last 250 observations as a 1-step-ahead VaR forecast. 25000 replications of data, using sample sizes $n = 250, 500, 1000$ and 2500, are simulated in each case. For each data set, the UC, CC, IND, DQ1, DQ4 and VQR tests are conducted. Further, the Bp11, Bp55 intervals are also calculated, as are the BFUQ, BFIND, BFCC, BFDQ1, BFDQ4 and BFVQ statistics. These are all calculated under the null, using the true $\text{VaR}_{t,\alpha}$ series, and then separately calculated under the alternative, using the estimated $\text{HS250}_{t,\alpha}$ series. $\alpha = 0.05, 0.01$ are used for the quantile levels.

So as to compare the methods on an equal footing we consider frequentist size and power as well as empirically adjusted size and size-adjusted power. This is standard practice when comparing frequentist tests, but is not standard for Bayesian methods. Whilst BF methods generally use BF = 1 as the threshold for a decision rule, there is no reason why that point should have a frequentist size equal to nominal (here 5%). To properly compare the sampling properties of all these tests, we thus consider both the unadjusted and adjusted size and power characteristics of each. Such will allow direct, fair and objective comparison of all methods in Sections 2 and 3 on an equitable basis.

Table 1 shows the empirical estimates for size and empirically adjusted size, across all the methods employed at $\alpha = 0.05$ for $n = 250, 500$. Also shown are the empirical 5% points for all methods, calculated via relevant sample percentiles across the 25000 replications of each test statistic under the null hypothesis. These are the thresholds used to calculate the empirically adjusted size and size-adjusted power below; i.e. adjusted size is the observed percentage of test statistics, across the 25000 replications, that are beyond the empirical 5% threshold, all calculated under the null hypothesis. Size-adjusted power is the observed percentage of test statistics beyond the same empirical threshold, when calculated under the incorrect HS VaR estimator. There is no reason why the point BF= 1 should be the 95th percentage point for the sampling distribution of any BF statistic, so size for the BF methods is not particularly relevant, but is reported as a reference for comparison. Adjusted size is, however, relevant to the comparison of all methods.

When $\alpha = 0.05$ and $n = 250$, only the VQR and DQ1 tests achieve close to a true nominal size, with $4.7\%, 5.5\%$ respectively, whilst the DQ4 is quite over-sized and the IND is quite under-sized; however the DQ1, DQ4, VQR, BFDQ1, BFDQ4 and BFVQ all achieve (very close to) correct adjusted sizes of exactly 5%. Tests whose empirical

Table 1: Size, empirical threshold and adjusted size for nominal 5% tests of 1-step-ahead quantile forecasts at $\alpha = 0.05$.

	n=250			n=500		
Method	Size	threshold	adj. size	Size	threshold	adj. size
UC	0.0623	4.040	0.0482	0.0548	3.888	0.0400
BFUC	0.0132	0.218	0.0437	0.0074	0.188	0.0400
Bp11	0.0403	(5,18)	0.0535	0.0512	(16,34)	0.0512
Bp55	0.0594	(5,18)	0.0535	0.0512	(16,34)	0.0512
IND	0.0179	2.808	0.0436	0.0337	3.581	0.0416
BFIND	0.0365	0.758	0.0496	0.0222	0.493	0.0498
CC	0.0422	5.226	0.0492	0.0400	5.751	0.0497
BFCC	0.0036	0.088	0.0460	0.0017	0.042	0.0487
DQ1	0.0551	8.024	0.0500	0.0471	7.666	0.050
BFDQ1	0.0018	0.00025	0.0500	0.0002	5.69×10^{-5}	0.050
DQ4	0.0669	13.833	0.0500	0.0544	12.877	0.050
BFDQ4	0.0005	1.574×10^{-7}	0.0500	0.00004	8.13×10^{-9}	0.050
BFLDQ1	0.0796	7.376	0.0500	0.4722	3.45×10^{7}	0.050
BFLDQ4	0.0134	1.276×10^{-5}	0.0500	0.1734	4.66×10^{13}	0.050
VQR	0.0470	5.774	0.0500	0.0559	6.363	0.050
BFVQ	0.1920	113.321	0.0500	0.1599	99.315	0.050

The correct, nominal thresholds are UC, IND 3.84; CC, VQ 5.99; DQ1 7.81; DQ4 12.59

distributions are discrete, with fixed values depending on the number of violations, i.e. UC, BFUC, Bp11, Bp55, cannot be so accurately corrected, as neither can the IND test. The threshold entries for Bp11 and Bp55 (5,18) indicate an approximate 95% (i.e. 94.65%) prediction interval for the number of violations under the null hypothesis. For $n = 500$ the Bp11, Bp55 methods have the closest to nominal size (5.1%), followed by DQ1 (4.7%); while for most methods the corrected size is exactly, or very close to, nominal, excepting UC, BFUC and IND.

Table 2 shows the empirical estimates for size and then adjusted size across all the methods employed at $\alpha = 0.05$ for $n = 1000, 2500$. Also shown are the empirical 5% points for the methods across the 25000 replications used to calculate the adjusted size and size-adjusted power.

At $n = 1000$, the Bp11, Bp55 are again closest to nominal size, with the VQR and DQ4 methods also achieving very close to nominal (5.1%); the IND test is well over-sized; however all methods achieve corrected sizes of close to or exactly equal to 5%. The threshold entries for Bp11 and Bp55 (37,63) indicate an approximate 95% prediction interval for the number of violations when $n = 1000$ under the null hypothesis. For $n = 2500$ the DQ4 has the closest to nominal size (5.02%) and for all methods the corrected size is very close to the nominal 5% level.

Table 3 shows the empirical estimates for power and size-adjusted power across all the methods employed at $\alpha = 0.05$. At $n = 250$ three methods stand out with power: BFVQ, DQ1 and DQ4; however we know from Table 1 that all of these are over-sized. When using the thresholds in Table 1 to calculate size-adjusted power the single stand-out method is clearly BFDQ1 with $\sim 43\%$; BFDQ4, DQ1 and DQ4 are next best with $\sim 35\%$, then BFVQ with 30%. To examine this result in more detail, consider Figure 1(a), showing two times the logarithm of the BFDQ1 and DQ1 statistics, plot against the number of violations from the incorrect HS 5% VaR estimator. As expected, both methods reject the HS estimator for very low and very high numbers of violations (e.g. Bp11 rejects outside of (5,18)). However, unlike the Bp11, Bp55, UC and BFUC methods, both BFDQ1 and DQ1 also have plenty of rejections inside the range (5,18): i.e. for violation series showing "significant" correlation. This explains the higher size-adjusted power achieved by DQ1 and BFDQ1 (and also DQ4, BFDQ4, BFVQ, CC, BFCC, etc) over the Bp11,

Table 2: Size, empirical threshold and adjusted size for nominal 5% tests of 1-step-ahead quantile forecasts at $\alpha = 0.05$.

	n=1000			n=2500		
Method	Size	threshold	adj. size	Size	threshold	adj. size
UC	0.0534	3.895	0.0534	0.0530	3.867	0.0530
BFUC	0.0043	0.130	0.0430	0.0024	0.071	0.0478
Bp11	0.0492	(37,63)	0.0492	0.0480	(104,146)	0.0480
Bp55	0.0492	(37,63)	0.0492	0.0544	(104,146)	0.0480
IND	0.0840	4.643	0.0487	0.0525	3.951	0.0498
BFIND	0.0168	0.421	0.0486	0.0123	0.320	0.0500
CC	0.0572	6.231	0.0500	0.0549	6.301	0.0500
BFCC	0.0007	0.023	0.0492	0.0004	0.010	0.0500
DQ1	0.0470	7.680	0.0500	0.0482	7.739	0.0500
BFDQ1	0.0000	1.724×10^{-5}	0.0500	0.0000	3.759×10^{-6}	0.0499
DQ4	0.0512	12.694	0.0500	0.0502	12.599	0.0500
BFDQ4	0.0000	7.025×10^{-10}	0.0500	0.0000	3.517×10^{-11}	0.0500
BFLDQ1	0.9071	1.418×10^{24}	0.0500	0.9993	1.402×10^{90}	0.0500
BFLDQ4	0.6962	3.467×10^{132}	0.0500	0.9996	exp(3884.7)	0.0500
VQR	0.0511	6.042	0.0500	0.0478	5.911	0.0500
BFVQ	0.1177	45.142	0.0500	0.0481	0.881	0.0500

The correct, nominal thresholds are UC, IND 3.84; CC, VQ 5.99; DQ1 7.81; DQ4 12.59

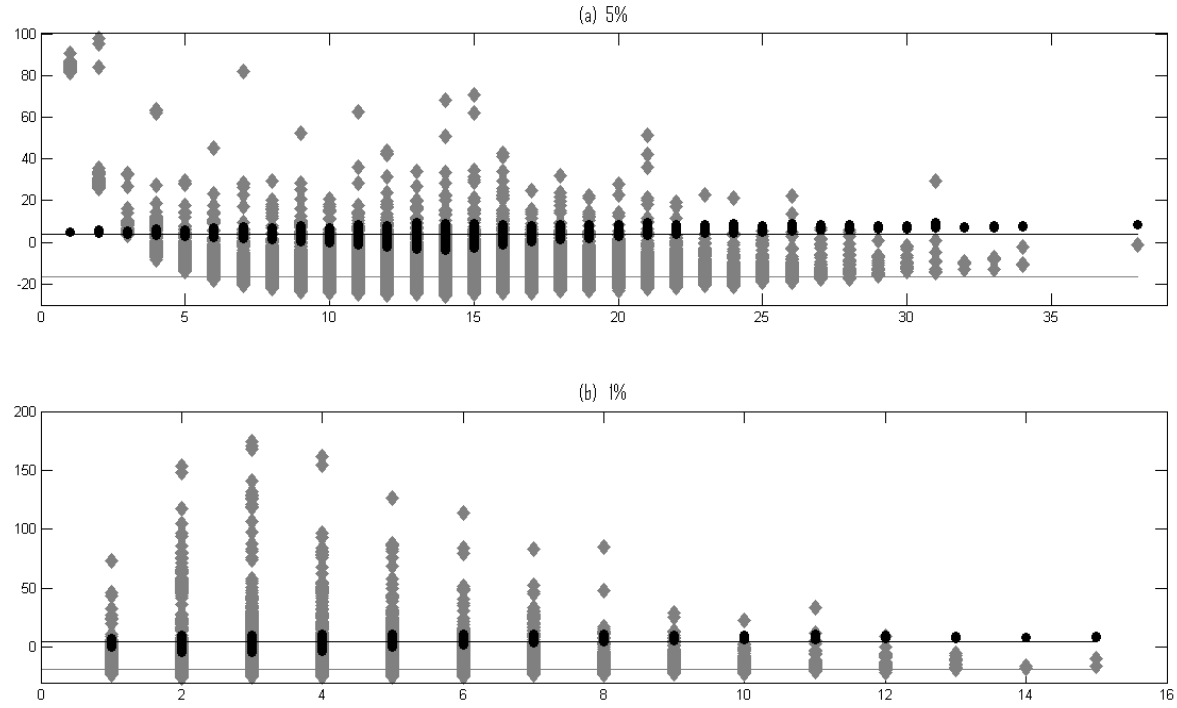

Figure 1: Two times the logarithm of the BFDQ1 (diamond) and DQ1 (circle) statistics against number of violations under the HS quantile estimator at $n = 250$; (a) 5% HS; (b) 1% HS. The horizontal lines are two times the logarithm of the empirical 5% points of the DQ1 (black) and BFDQ1 (grey) statistics.

Bp55, UC and BFUC methods. The extra power of BFDQ1 over DQ1 is achieved through rejecting more when the number of violations is comparatively low, specifically for 3-14 violations; the DQ1 has more rejections than BFDQ1 when 15-28 violations are observed, but the differences in rejection frequencies here are much smaller than they are for 3-14 violations; as shown in Figure 2(a). Thus, at n=250 and 5% VaR forecasting, the BFDQ1 appears to have much higher power than DQ1 at detecting correlation in the violation series, especially when the number of violations is small (approximately 3-14 violations), contributing to an overall higher size adjusted power in Table 3 when $n = 250$.

At $n = 500$, the two methods with highest size-adjusted power for 5% quantile forecasting are: BFDQ4 and BFDQ1 with $\sim 50\%$, followed by DQ4, DQ1 with $49\%, 46\%$ and then BFVQ with 34%. The BFDQ4 is only marginally preferred here. The BFDQ4 (BFDQ1) test rejects more than the DQ4 (DQ1) test when the number of violations is between 8 and 27 (12 and 28), whilst the DQ4 rejects more for 28-46 (DQ1 for 29-42) violations: again the BFDQ statistics have slightly more power to detect correlated violations when there are lower numbers of violations, compared to the DQ tests; however, in this case things even out so that the Bayesian and frequentist DQ method's size-adjusted powers are close to comparable in each case. For $n = 1000$, again the BFDQ4 is marginally the highest size-adjusted power, followed closely by the DQ4, then with BFDQ1 marginally out-performing the DQ1. The BFDQ4 (BFDQ1) test rejects more than the DQ4 (DQ1) test when the number of violations is between 33 and 52 (33 and 55), whilst the DQ4 rejects more for 53-73 (DQ1 for 56-71) violations: again the BFDQ statistics have slightly more power to detect correlation at lower numbers of violations, compared to the DQ tests, with the result reversed for higher violation numbers. This is illustrated by Figure 3(a), comparing DQ1, BFDQ1 and BFLDQ tests in terms of rejection rates against number of violations when $n = 1000$. Clearly, the BFLDQ methods are not at all powerful at 5%, compared to the BFDQ and DQ methods.For $n = 2500$, similar results are obtained, but with a marked increase in size-adjusted power for the BFDQ, DQ, BFVQ and VQR methods. Again the BFDQ4 methods marginally performs the best with very high size-adjusted power of 97.4%, closely matched by DQ4 and then followed closely by BFDQ1, DQ1 and BFVQ, which marginally out-performs the VQR.

In summary for 5% quantile forecasting: the BFDQ1 method clearly out-performed

Table 3: Power and two estimates of size-adjusted power for nominal 5% tests of 1-step-ahead quantile forecasts at $\alpha = 0.05$.

	n=250		n=500		n=1000		n=2500	
Method	Power	Size-adj.	Power	Size-adj.	Power	Size-adj.	Power	Size-adj.
UC	0.159	0.130	0.057	0.051	0.025	0.025	0.026	0.026
BFUC	0.068	0.134	0.012	0.051	0.002	0.024	0.0003	0.026
Bp11	0.135	0.152	0.068	0.068	0.035	0.035	0.029	0.029
Bp55	0.160	0.152	0.068	0.068	0.035	0.035	0.029	0.029
IND	0.087	0.146	0.159	0.169	0.260	0.215	0.510	0.500
BFIND	0.143	0.174	0.163	0.258	0.228	0.346	0.384	0.559
CC	0.161	0.182	0.132	0.147	0.191	0.188	0.424	0.417
BFCC	0.050	0.194	0.036	0.194	0.035	0.255	0.087	0.459
DQ1	0.364	0.349	0.447	0.456	0.625	0.634	0.945	0.946
BFDQ1	0.039	0.426	0.020	0.506	0.028	0.673	0.110	0.925
DQ4	0.387	0.343	0.505	0.492	0.709	0.706	0.972	0.972
BFDQ4	0.019	0.360	0.009	0.500	0.013	0.710	0.063	0.973
BFLDQ1	0.103	0.066	0.465	0.043	0.981	0.031	1.000	0.025
BFLDQ4	0.051	0.210	0.391	0.136	0.963	0.175	1.000	0.338
VQR	0.159	0.165	0.272	0.260	0.477	0.473	0.873	0.877
BFVQ	0.565	0.309	0.566	0.336	0.649	0.431	0.920	0.926

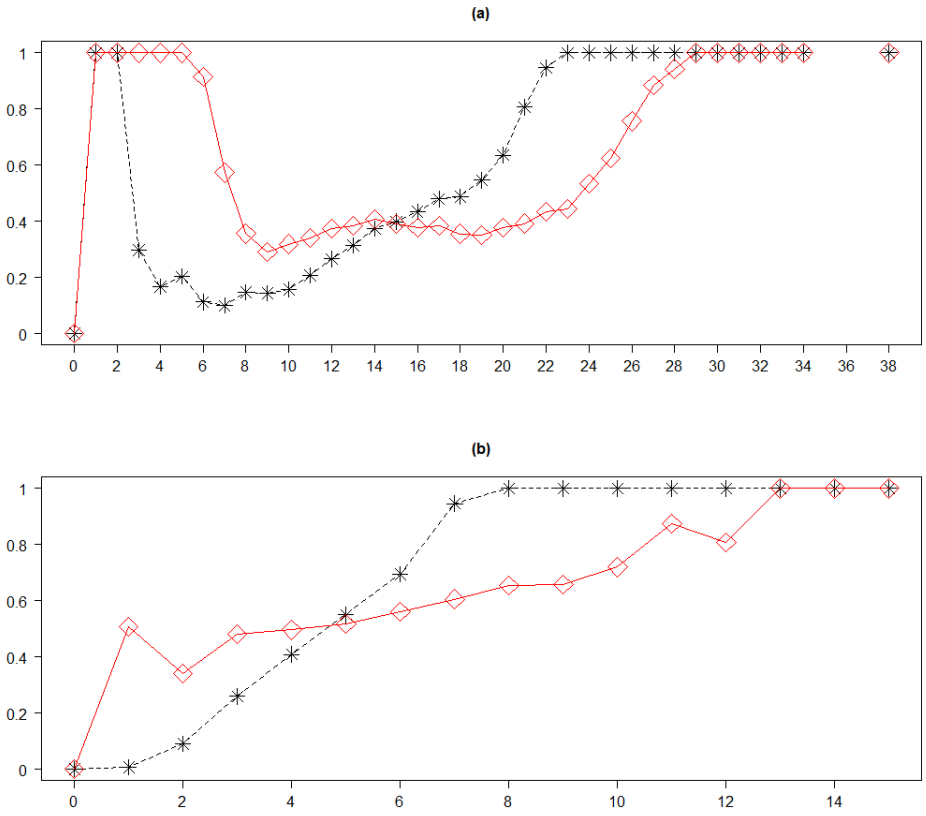

Figure 2: Rates of rejection for the BFDQ1 (circle) and DQ1 (diamond) statistics against number of violations under the HS quantile estimator at $n = 250$; (a) 5% HS; (b) 1% HS.

all other methods when $n = 250$ regarding size-adjusted power. For $n > 250$, the BFDQ4 method marginally out-performed all other methods on this criterion, closely followed by the DQ4, BFDQ1 and DQ1 methods, usually in that order. The out-performance in size-adjusted power is attributable to more accurate detection of autocorrelation in violations from the HS estimator, when the observed violation numbers were comparatively small. Further, the BFVQ test had higher size-adjusted power than the VQR test, except when $n = 1000$. The BFLDQ methods performed quite poorly in comparison.

Table 4 shows the empirical estimates for size and then adjusted size across all the methods employed at $\alpha = 0.01$ for $n = 250, 500$. Also shown are the empirical 5% points for the methods. Again, empirical size for the BF methods is only reported as a reference for comparison.

At $n = 250$, the DQ1 test achieves closest to the true nominal size, with 5.4%, followed by BP11 (4.2%); whilst the UC, DQ4, Bp55 and VQR are well over-sized, and the IND and CC are well under-sized. Only the VQR and BFVQ tests achieve correct adjusted sizes of 5%; whilst the UC, IND, BFIND, CC and BFCC adjusted sizes are not close to nominal. The "thresholds" for Bp11 and Bp55 of (0,5) indicate an approximate 95% prediction interval for the number of violations under the null. For $n = 500$ VQR (3.9%) and Bp11 (3.7%) have the closest to nominal sizes, but most methods have a corrected size that is reasonably close to nominal, except UC and CC.

Table 5 shows the empirical estimates for size and then adjusted size across all the methods employed at $\alpha = 0.01$ for $n = 1000, 2500$. Also shown are the empirical 5% points for the methods across the 25000 replications used to calculate the adjusted size and size-adjusted power.

At $n = 1000$, only the UC and Bp55 methods achieve close to the true nominal size ($\sim$ 5.5%), whilst the DQ1, DQ4 (and BFVQ) are well over-sized; however all methods achieve corrected sizes of close or exactly equal to 5%, except BFUC and CC. The threshold entries for Bp11 and Bp55 (5,16) indicate an approximate 95% prediction interval for the number of violations when $n = 1000$ under the null. For $n = 2500$ the VQR test has the closest to nominal size (5.1%), whilst most methods achieve an adjusted size close to nominal, except UC and BFUC.

Table 6 shows the empirical estimates for power and size-adjusted power across all

Table 4: Size, empirical threshold and adjusted size for nominal 5% tests of 1-step-ahead quantile forecasts at $\alpha = 0.01$.

	n=250			n=500		
Method	Size	threshold	adj. size	Size	threshold	adj. size
UC	0.095	5.025	0.0154	0.071	4.813	0.0200
BFUC	0.004	0.060	0.0422	0.002	0.060	0.0378
Bp11	0.042	(0,5)	0.0399	0.037	(1,9)	0.0374
Bp55	0.125	(0,5)	0.0399	0.073	(1,9)	0.0374
IND	0.013	0.296	0.0368	0.016	0.991	0.0500
BFIND	0.023	1.000	0.0216	0.033	0.715	0.0464
CC	0.007	5.025	0.0310	0.017	4.817	0.0307
BFCC	0.0008	0.049	0.0367	0.001	0.030	0.0445
DQ1	0.054	8.924	0.0460	0.073	12.135	0.0497
BFDQ1	0.007	8.391×10^{-5}	0.0460	0.007	0.0030	0.0500
DQ4	0.096	26.250	0.0460	0.168	21.172	0.0496
BFDQ4	0.008	1.577×10^{-6}	0.0458	0.010	1.173×10^{-5}	0.0497
BFLDQ1	0.001	0.0011	0.0460	0.007	0.0039	0.0497
BFLDQ4	0.000	4.692×10^{-8}	0.0458	0.000	6.532×10^{-8}	0.0496
VQR	0.070	8.016	0.0500	0.039	5.037	0.0500
BFVQ	0.569	1.845×10^{10}	0.0500	0.508	9.199×10^{10}	0.0500

The correct, nominal thresholds are UC, IND 3.84; CC, VQ 5.99; DQ1 7.81; DQ4 12.59

Table 5: Size, empirical threshold and adjusted size for nominal 5% tests of 1-step-ahead quantile forecasts at $\alpha = 0.01$.

	n=1000			n=2500		
Method	Size	threshold	adj. size	Size	threshold	adj. size
UC	0.0548	4.091	0.0548	0.0430	3.752	0.0584
BFUC	0.0024	0.054	0.0357	0.0014	0.036	0.0430
Bp11	0.0362	(5,16)	0.0546	0.0529	(16,34)	0.0529
Bp55	0.0546	(5,16)	0.0546	0.0424	(16,34)	0.0529
IND	0.0192	2.290	0.0495	0.0171	1.888	0.0462
BFIND	0.0275	0.611	0.0495	0.0154	0.305	0.0456
CC	0.0264	4.738	0.0383	0.0284	4.897	0.0485
BFCC	0.0012	0.011	0.0447	0.0002	0.005	0.0499
DQ1	0.0822	9.626	0.0500	0.0571	8.395	0.0500
BFDQ1	0.0073	1.054×10^{-4}	0.0500	0.0008	9.610×10^{-6}	0.0500
DQ4	0.1057	17.103	0.0500	0.0996	15.568	0.0500
BFDQ4	0.0055	7.585×10^{-7}	0.0500	0.0004	3.473×10^{-9}	0.0500
BFLDQ1	0.0337	0.2060	0.0500	0.2183	1.343×10^{6}	0.0500
BFLDQ4	0.0001	1.099×10^{-7}	0.0500	0.0056	6.986×10^{-7}	0.0500
VQR	0.0387	5.1088	0.0500	0.0508	6.062	0.0500
BFVQ	0.4561	5.211×10^{10}	0.0500	0.3288	3.294×10^{9}	0.0500

The correct, nominal thresholds are UC, IND 3.84; CC, VQ 5.99; DQ1 7.81; DQ4 12.59

the methods at $\alpha = 0.01$. At $n = 250$ one method stands out with power: BFVQ; however we know from Table 4 that it is well over-sized. When using the thresholds in Table 4 to calculate size-adjusted power the single stand-out method is clearly BFDQ1 with 45%; the logistic regression based BFLDQ4 is next best with 34%, followed by DQ1 with 31%, then BFLDQ1, DQ4, BFDQ4 and BFVQ methods ($\sim 26\%$). To examine power in more detail, consider Figure 1(b), showing two times the logarithm of the BFDQ1 and DQ1 statistics, plot against the number of violations from the incorrect HS 1% VaR estimator. As is logical, with an expected number of only 2.5 violations under the null, both methods always reject the HS estimator only for very high numbers of violations (13 and above; e.g. Bp11 always rejects above 5). The out-performance of BFDQ1 is again due to a much higher rate of model rejection for low violation numbers, here 1-4, as shown in Figure 2(b) (comparing BFDQ1 and DQ1 in terms of rejection rates against number of violations), all of which occur highly frequently at $n = 250$ and 1% HS forecasting. The DQ1 has higher power than BFDQ1 at 5-12 violations, but these are far less likely to occur in this case.

However, the situation is very different for $n > 250$ here. For $n = 500$ and $n = 1000$, the stand-out method with highest size-adjusted power in each case is the BFLDQ4 (50% and 68% respectively). In each case the DQ4 (43%, 60%) and DQ1, BFVQ methods (36%, 57%) are next best. In this case, the BFDQ1 and BFDQ4 statistics still out-performed the DQ1 and DQ4 respectively, only when e.g. 1-4 violations were observed for $n = 500$, but these are not very likely outcomes when $n = 500$ at 1% forecasting, and the out-performance of the DQ statistics for the more frequently occurring violation numbers of 5-12 was enough for both DQ statistics to overall clearly out-perform both BFDQ statistics here. The out-performance of BFLDQ4, compared to DQ and BFDQ methods at $n = 500, 1000$ is fairly uniform across the observed violation numbers, as illustrated by Figure 3(b), comparing DQ1, BFDQ1 and BFLDQ tests in terms of rejection rates against number of violations. Similar results occurred for $n = 2500$, in each case the DQ statistics clearly out-performed their BFDQ counterparts, but both were out-performed by BFLDQ4. However, at $n = 2500$ the BFVQ test recorded clearly the highest size-adjusted power with 95%, followed by BFLDQ4 (93%), DQ1, DQ4 with $\sim$ 91% and BFDQ1, BFDQ4 with $\sim 80, 84\%$.

Overall, for 1% quantile forecasting, the results are mixed, though always favour

Table 6: Power and size-adjusted power for nominal 5% tests of 1-step-ahead quantile forecasts at $\alpha = 0.01$.

	n=250		n=500		n=1000		n=2500	
Method	Power	Size-adj.	Power	Size-adj.	Power	Size-adj.	Power	Size-adj.
UC	0.150	0.082	0.135	0.070	0.143	0.143	0.296	0.296
BFUC	0.042	0.155	0.016	0.129	0.015	0.143	0.020	0.296
Bp11	0.150	0.150	0.125	0.125	0.139	0.139	0.385	0.385
Bp55	0.221	0.150	0.131	0.125	0.139	0.139	0.301	0.385
IND	0.066	0.184	0.084	0.257	0.140	0.206	0.256	0.456
BFIND	0.112	0.112	0.157	0.193	0.150	0.205	0.233	0.426
CC	0.075	0.161	0.093	0.139	0.186	0.220	0.385	0.489
BFCC	0.028	0.184	0.029	0.206	0.036	0.243	0.068	0.537
DQ1	0.358	0.330	0.466	0.359	0.667	0.570	0.932	0.916
BFDQ1	0.060	0.457	0.054	0.161	0.057	0.369	0.107	0.841
DQ4	0.410	0.301	0.630	0.428	0.755	0.604	0.952	0.913
BFDQ4	0.070	0.276	0.075	0.259	0.070	0.351	0.113	0.802
BFLDQ1	0.025	0.278	0.066	0.328	0.274	0.352	0.825	0.475
BFLDQ4	0.001	0.342	0.005	0.498	0.031	0.679	0.451	0.933
VQR	0.097	0.073	0.064	0.075	0.149	0.172	0.571	0.568
BFVQ	0.805	0.261	0.893	0.346	0.950	0.575	0.994	0.946

The correct, nominal thresholds are UC, IND 3.84; CC, VQ 5.99; DQ1 7.81; DQ4 12.59

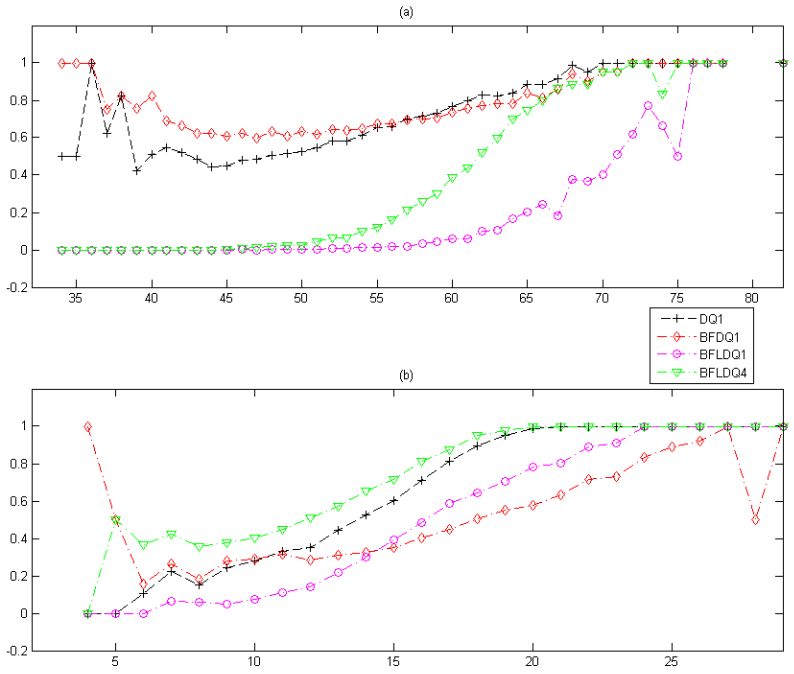

Figure 3: Rates of rejection for the DQ1 (cross), BFDQ1 (diamond), BFLDQ1 (circle) and BFLDQ4 (triangle) statistics, against number of violations under the HS quantile estimator at $n = 1000$; (a) 5% HS; (b) 1% HS.

a Bayesian method. For $n = 250$, the BFDQ1 method is clearly favoured. At $n = 500, 1000$ however the BFLDQ4 method is highly favoured. Finally, at $n = 2500$, the BFVQ method is marginally favoured regarding size-adjusted power, closely followed by BFLDQ4. Further, the BFVQ method always had higher size-adjusted power than the VQR test, as discussed next.

5 Discussion

When detecting the incorrect HS estimator of 1% and 5% quantiles using a range of competing tests/methods, fairly similar stories can be told at each quantile level, but with some important differences. First, the DQ and BFDQ methods were almost always prevalent at or near the top of the methods regarding size-adjusted power (except BFDQ1 at $n = 500$ ranking 7th). For the BFLDQ methods, this statement only holds for 1% forecasting; these methods were usually the worst for size-adjusted power at 5% forecasting. On the contrary, the UC, BFUC, IND, BFIND, CC, BFCC, BFLDQ1 and VQR methods almost always performed towards the bottom on this aspect (except VQR at $n = 2500$ ranking 6th for both $\alpha = 0.01, 0.05$). The relatively lower (size-adjusted) powers observed for the UC, BFUC, IND, BFIND, CC and BFCC methods are not surprising: it is well known that DQ and VQR (and hence BFDQ) methods use more information and outperform the frequentist versions of these tests, see e.g. Berkowitz et al. (2011). The very low power observed for the UC, BFUC, Bp11 and Bp55 tests for 5% VaR forecasting also make sense since, even in small samples, the 5% HS VaR estimator will give typically give close to the correct, nominal violation rate; in fact we observed a fixed ratio of average number of violations from the HS estimator, compared to that expected under the null, of 1.065 (i.e. only 6.5% more violations on average under the HS estimator), that was consistent across n. This ratio increased to 1.34 for the 1% HS VaR estimator, again consistent across n. The IND, BFIND, CC and BFCC tests easily out-performed the UC test on size-adjusted power, for both 5% and 1% VaR forecasting, because the HS estimator tends to generate highly correlated violations (it employs heavily overlapping data periods to generate successive VaR estimates) that are nevertheless close to correct on average (typically only 6% higher in violation rate at 5% VaR and 34% higher at 1%

VaR). However, for 1% VaR forecasting the UC, BFUC tests were more powerful than at 5% VaR forecasting, because of the higher discrepancy in average violation rate ratios: 34% more violations is easier to detect than 6% more, on average. In addition, the ratio of the range of the numbers of violations under the 5% HS estimator, compared to that under the true VaR series, decreases with n, being 1.14 for $n = 250$ but only 0.64 for $n = 2500$; this is partly because the HS estimator "follows" the data pattern, in a non-parametric manner, and so comparatively extreme numbers of violations are highly unlikely to occur, and become less likely as n increases: thus as the UC, BFUC, Bp11 and Bp55 tests can only reject for extreme numbers of violations, they have lower power as n increases for 5% VaR forecasting, compared to that for 1% forecasting.

The reported performance of the VQR test in terms of size-adjusted power is slightly worse than the results in Gaglianone et al. (2011), though nearly comparable. However, the performance of the DQ methods here is much better than that in Gaglianone et al. (2011): we speculate that is because we included the VaR_{t+1} forecast in the design matrix for the DQ tests, whilst Gaglianone et al. (2011) only included lagged "hits" (and only 1 lag).

The relatively poor performance of the VQR, compared to the BFVQ, test bears more examination. Figure 4 shows two times the logarithm of the BFVQ (upper) and VQ (lower) test statistics, plot against the number of violations, under the null hypothesis (black circles) and also under the HS quantile estimator (grey diamonds), at $n = 1000$. Also shown are two times the logarithm of the empirical 5% points of the BFVQ (upper) and VQR (lower) statistics; points above these lines represent rejections of the null hypothesis. It is immediately apparent that slightly more violations tend to occur under the HS estimator than under the null, the latter having a distribution shifted to the right compared to the former (the ratio of means is 1.34 as mentioned previously). Further, the BFVQ statistic has a clearly distinguished distribution of values, typically higher, under the HS estimator (grey diamonds) compared to that under the null; this leads to the observed 57.5% size-adjusted power of the BFVQ method. On the contrary, the VQR test statistic does not have a clearly distinguished distribution of values under the HS estimator compared to that the null, leading to its very low size adjusted power (17.4%). Similar illustrations, not shown, occur at $n = 250, 500, 2500$.

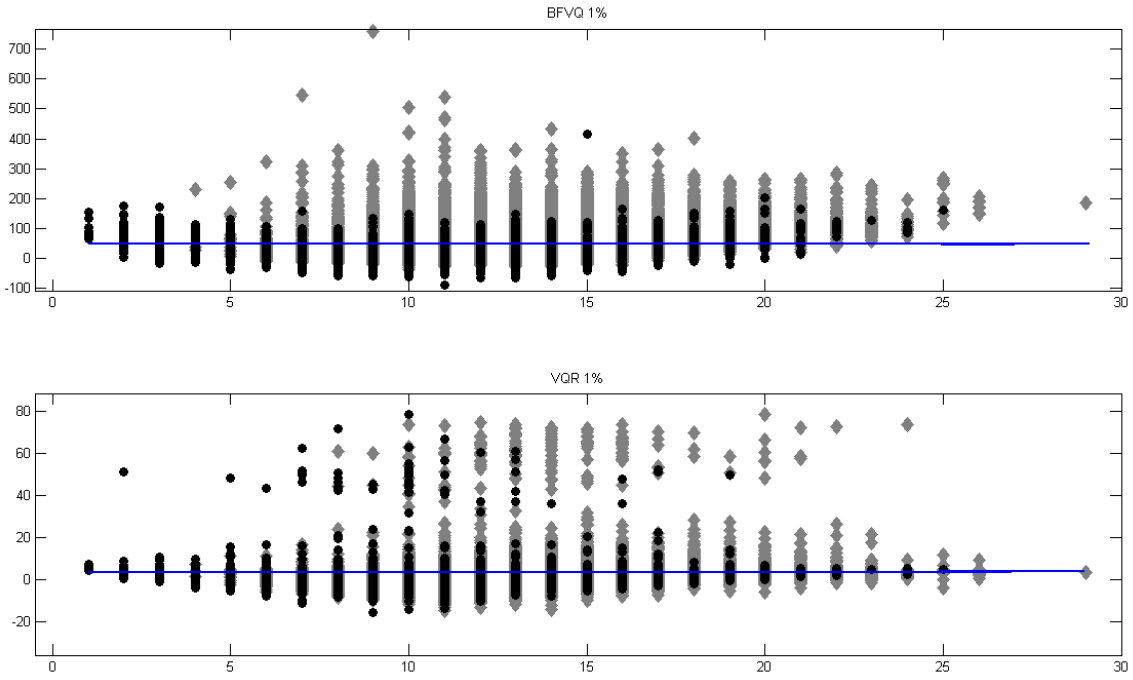

Figure 4: Two times the logarithm of the (a) BFVQ and (b) VQ test statistics, plot against the number of violations, under the null hypothesis (black circles) and the HS quantile estimator (grey triangles) at $n = 1000$. The horizontal lines are two times the logarithm of the empirical 5% points of the BFVQ (a) and VQR (b) statistics.

When comparing the Bayesian version of each test with its frequentist counterpart (e.g. BFDQ1 vs DQ1, etc), at the 5% quantile level the results are very clear: at each sample size the Bayesian version of each test had higher size-adjusted power, often only marginally but sometimes much, much higher, than its frequentist competitor, for all the tests considered (except VQR, BFVQ at $n = 1000$). This is a very strong and clear result in favour of the Bayesian method at the 5% quantile level.

For 1% quantile forecasting, the results are not so consistent: the UC and BFUC have virtually identical size-adjusted powers over n; though Bp11 and Bp55 are marginally higher in each case. The IND and CC tests had lower size-adjusted power than the BFIND and BFCC at all sample sizes; the VQR had much lower size-adjusted power than the BFVQ for all n; but the DQ1 had higher size-adjusted power than the BFDQ1 for $n > 250$, but lower for $n = 250$. Finally the DQ4 had higher size-adjusted power than the BFDQ4 for all sample sizes. As noted, the results for 1% quantile forecasting are mixed.

6 Empirical study

We briefly report the results of a large empirical study here. Seven daily financial time series: prices, exchange rates or financial indices, are considered, in each case converting these to daily percentage log returns. The seven series are: the US S&P500 index, the AORD, FTSE100 and Hang Seng indices, the AU US exchange rates, the EU US exchange rates and IBM asset prices. The initial sample period is specifically from January 2, 1998 to December 15, 2005, approximately 2000 days in each case. The forecast period is December 16, 2005 to January 15, 2010, covering close to 1000 trading days in each market, and including the well-known global financial crisis (GFC) period.

One-step-ahead forecasts of VaR at 5% and 1% quantile levels are estimated under a range of competing models and methods, for each day in the forecast period. Forecasts for each of four types of heteroskedastic model: the GARCH of Bollerslev (1986), the GJR-GARCH of Glosten et al (1993) the Threshold (T-)GARCH in Chen and So (2006) and a smooth transition (ST-)GARCH as in Gerlach and Chen (2008) are estimated employing the MCMC methods of Chen et al. (2012b). Each of these specifications is estimated under five types of error distribution: Gaussian, Student-t, the skewed Student-t of Hansen

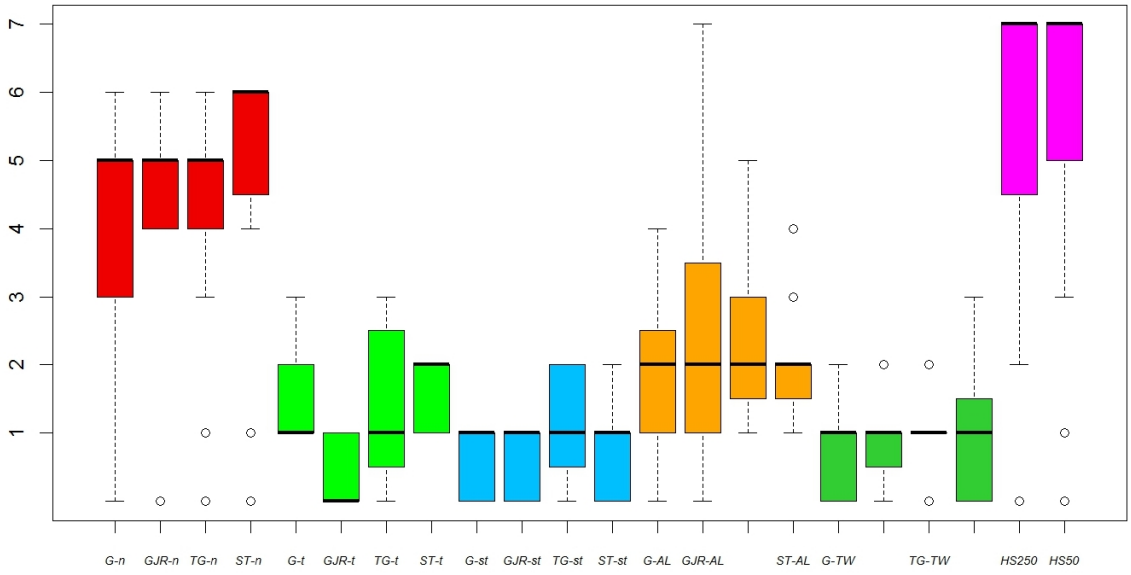

Figure 5: Boxplots of rejection counts for each 1% VaR model across the seven asset series. Each boxplot represents 11 counts of rejections, out of 7 series.

(1994), the Asymmetric Laplace (AL) of Chen et al (2012c) and the Two-sided Weibull (TW) of Chen and Gerlach (2013). This gives 20 models generating VaR forecasts at 5% and 1% quantile levels for 1000 days. Also considered are the non-parametric 50 day and 250 day sample percentile HS methods, thus giving a total of 22 models or methods. Estimation results are not shown to save space, since only the test results are directly relevant to this paper; it is expected that most models and methods will be rejected since the data includes the GFC period, where that outcome is common; but models and methods that can better captured highly changing volatility and fat-tailed returns will be rejected the least, across the seven series.

Tables 7 and 8 show the number of series, out of 7, that each model or method of VaR estimation was rejected in, using the UC, BFUC, Bp11, DQ1, BFDQ1, BFLDQ1, DQ4, BFDQ4, BFLDQ4, VQR and BFVQ tests for 5% and 1% VaR forecasting, respectively. The tests were conducted at the 5% level using the empirical cut-offs in Tables 1-5 above. Results shaded red indicate the most rejections for each model, boxes indicate the fewest rejections.

Alternatively, the results from these tables can be evaluated based on the model/method classifications which are illustrated in Figures 5 and 6. The figures illustrate that models

Table 7: Number of rejections for each model across 7 series for 5% VaR forecasting.

Method	UC	BFUC	Bp11	DQ1	BFDQ1	BFLDQ1	DQ4	BFDQ4	BFLDQ4	VQ	BFVQ
G-n	[0]	[0]	2	[0]	[0]	5	3	2	5	[0]	[0]
GJR-n	3	3	5	2	[0]	4	3	2	4	1	2
TG-n	3	3	4	3	1	6	3	1	6	2	2
ST-n	4	4	4	3	1	6	3	2	5	1	2
G-t	1	1	3	1	1	2	4	4	3	2	2
GJR-t	1	1	3	1	1	2	2	2	2	1	1
TG-t	2	2	3	1	1	3	2	2	2	2	1
ST-t	2	2	3	1	1	4	2	2	2	3	1
G-SKT	[0]	[0]	1	[0]	[0]	5	3	1	6	[0]	[0]
GJR-SKT	2	2	3	1	[0]	5	2	1	3	1	[0]
TG-SKT	2	2	3	2	[0]	6	1	[0]	5	2	1
ST-SKT	1	1	3	1	[0]	6	2	[0]	5	1	1
G-AL	[0]	[0]	2	[0]	[0]	3	2	2	2	1	[0]
GJR-AL	1	1	3	1	1	2	3	2	1	1	1
TG-AL	1	1	2	[0]	1	2	2	1	2	1	1
ST-AL	1	1	1	[0]	[0]	4	2	3	3	1	[0]
G-TW	[0]	[0]	1	[0]	[0]	4	3	2	5	[0]	[0]
GJR-TW	1	1	2	[0]	[0]	4	1	2	3	[0]	[0]
TG-TW	[0]	[0]	2	[0]	[0]	6	1	[0]	5	[0]	[0]
ST-TW	1	1	3	1	[0]	6	3	[0]	4	1	[0]
HS250	7	7	7	[0]	7	1	6	7	[0]	[0]	7
HS50	7	7	7	1	7	6	5	7	[0]	7	7

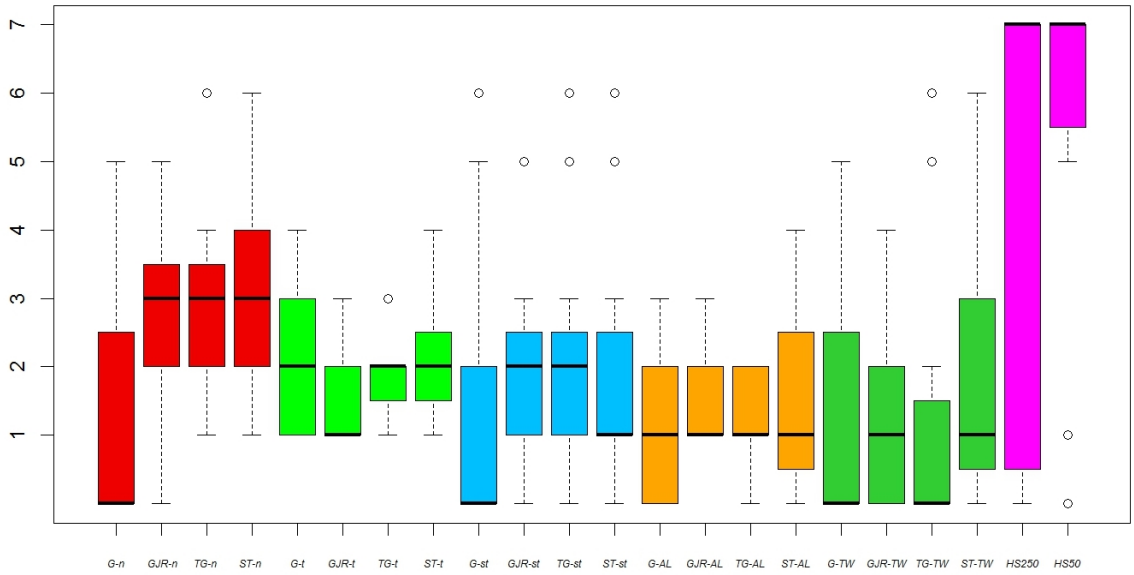

Figure 6: Boxplots of rejection counts for each 5% VaR model across the seven asset series. Each boxplot represents 11 counts of rejections, out of 7 series.

with Gaussian errors and HS methods are rejected the most, at both 5% and 1% VaR forecasting, and in all or most series. For 5% VaR forecasting, Student-t and Skewed-t error models have fewer rejections (improving on Gaussian errors), whilst models with AL and TW errors are generally rejected the least (except by the BFLDQ1 test) and are hence the most accurate 5% VaR forecasters in this set of models for this forecast time period. For 1% VaR forecasting, again Student-t and skewed-t error models have fewer rejections (improving on Gaussian errors), but here skewed-t seems preferable. Models with skewed-t or TW errors are generally rejected the least by all tests, and are hence the most accurate 1% VaR forecasters in this set of models for this forecast time period.

Comparing Bayesian and frequentist method results: the frequentist and Bayesian analogue pairs of tests mostly reject each model approximately the same number of times over the seven series at each quantile level. The UC and BFUC tests agree almost perfectly, though Bp11 mostly has 1-2 more rejections for each model, reflecting its slightly higher power. At 5% forecasting, the VQ and BFVQ mostly agree, except on model HS250 where the BFUC rejects it in all seven markets and the VQR never rejects it. However, both DQ tests tend to reject each model 1 or 2 more times than their BFDQ counterpart, except again for HS250 and HS50, where the BFDQ test reject more times

than their corresponding DQ test. Overall, however, both the BFLDQ1 and BFLDQ4 tests have mostly rejected the highest number for each model; again an exception is the HS250 and HS50 models. Similar comments apply to the 1% VaR test results in Table 8. Here again the BFLDQ methods generally reject the most for each model, though the DQ methods are also quite comparatively powerful in this respect.

7 Conclusion

Bayesian methods for assessing and testing forecast accuracy for dynamic quantile forecasts are developed. In a simulation study, at both $\alpha = 0.01, 0.05$ quantile levels, the corresponding Bayesian method in most cases had higher size-adjusted power than its competing frequentist analogue. Results for 5% VaR forecasting favoured the BF dynamic quantile (BFDQ) methods, while results for 1% VaR forecasting were more mixed, with the BFDQ1 favoured at $n = 250$, BFLDQ4 favoured for $n = 500, 1000$ and the BFVQ favoured at the largest sample size. The proposed BFVQ method was close to uniformly more powerful than the VQR test. These results suggest that Bayesian methods have much to offer for quantile forecast assessment and testing, compared to existing frequentist tests.

Acknowledgement The authors thank Mr Matt Read for some preliminary investigations on the BFUC, BFIND and BFCC statistics that appeared in his Honours thesis, Read (2011), and Declan Walpole for coding up the VQR test.

Table 8: Number of rejections for each model across 7 series for 1% VaR forecasting.

Model	UC	BFUC	Bp11	DQ1	BFDQ1	BFLDQ1	DQ4	BFDQ4	BFLDQ4	VQ	BFVQ
G-n	5	5	5	5	[0]	6	5	1	5	2	4
GJR-n	5	5	5	4	[0]	6	5	[0]	5	4	5
TG-n	5	5	5	6	1	6	5	[0]	5	3	5
ST-n	6	6	5	6	1	6	6	[0]	6	4	6
G-t	1	2	2	1	1	1	3	3	2	1	1
GJR-t	[0]	1	1	[0]	[0]	[0]	[0]	1	[0]	1	1
TG-t	3	3	3	[0]	1	[0]	[0]	1	1	1	2
ST-t	2	2	2	2	1	1	1	1	1	2	2
G-SKT	[0]	1	1	1	[0]	1	1	[0]	1	[0]	1
GJR-SKT	[0]	1	1	1	[0]	1	[0]	[0]	[0]	1	1
TG-SKT	2	2	2	1	[0]	1	[0]	1	1	[0]	2
ST-SKT	1	1	2	1	[0]	1	[0]	[0]	1	[0]	2
G-AL	1	1	2	1	2	[0]	2	4	2	3	3
GJR-AL	3	3	4	1	2	1	[0]	1	2	4	7
TG-AL	2	2	3	1	3	1	1	2	2	4	5
ST-AL	2	2	3	1	2	1	1	2	2	2	4
G-TW	[0]	[0]	2	1	[0]	1	[0]	[0]	1	1	1
GJR-TW	1	1	2	[0]	1	[0]	[0]	1	1	1	2
TG-TW	1	1	2	1	[0]	1	1	2	1	1	1
ST-TW	[0]	[0]	2	1	[0]	1	3	2	1	[0]	1
HS250	7	7	7	[0]	3	6	7	6	7	2	7
HS50	7	7	7	[0]	7	1	7	3	7	7	7

References

Berkowitz, J., Christoffersen, P., and Pelletier, D. (2011), "Evaluating Value-at-Risk Models with Desk-Level Data," *Management Science*, **57**, 2213-2227.

Bollerslev, T. (1986). "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, **31**, 307-327.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001) "Interval Estimation for a Binomial Proportion," *Statistical Science*, **16**, 101-117.

Chen, C. W. S., Gerlach, R., Hwang, R. B. K., and McAleer, M. (2012a), "Forecasting Value-at-Risk using nonlinear regression quantiles and the intra-day range," *International Journal of Forecasting*, **28**, 557-574.

Chen, C. W. S., Gerlach, R., Lin, E. M. H., and Lee, W.C.W. (2012b), "Bayesian forecasting for financial risk management, pre and post the global financial crisis," *Journal of Forecasting*, **31**, 661-687.

Chen, C. W. S., and So, M. K. P. (2006). "On a threshold heteroscedastic model," *International Journal of Forecasting*, **22**, 73-89.

Chen, Q., and Gerlach, R. (2013). "The two-sided Weibull distribution and forecasting financial tail risk," *International Journal of Forecasting*, **29**, 527-540.

Chen, Q., Gerlach, R. and Lu, Z. (2012c), "Bayesian Value-at-Risk and expected shortfall forecasting via the asymmetric Laplace distribution," *Computational Statistics & Data Analysis*, 1st Issue of Annals of Computational and Financial Econometrics, **56**, 3498-3516.

Christoffersen, P. F. (1998), "Evaluating interval forecasts," *International Economic Review*, **39**, 841-862.

Engle R. F., and Manganelli S. (2004), "CAViaR: Conditional autoregressive value at risk by regression quantiles," *Journal of Business and Economic Statistic*, **22**, 367-381.

Gaglianone, W. P., Lima, L. R., Linton, O., and Smith, D. R. (2011), "Evaluating Value-at-Risk models via quantile regression," *Journal of Business & Economic Statistics*, **29**, 150-160.

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

Gerlach, R., and Chen, C. W. S. (2008), "Bayesian inference and model comparison for asymmetric smooth transition heteroskedastic models," *Statistics and Computing*, **18**, 391-408.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). "On the relation between the expected value and the volatility of the nominal excess return on stocks," *Journal of Finance*, **487**, 1779-1801.

Hansen, B., (1994). "Autoregressive conditional density estimation," *International Economic Review*. **35**, 705 - 30.

Hoogerheide, L. F., and van Dijk, H.K. (2010), "Bayesian forecasting of Value at Risk and expected shortfall using adaptive importance sampling", *International Journal of Forecasting*, **26**, 231-247.

Kass, R. E. , and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, **90**, 773 - 795.

Koenker, R., and Machado, J. A. F. (1999) "Goodness of fit and related inference for quantile regression," *Journal of the American Statistical Association*, **94**, 1296-1310.

Kupiec, P. (1995) "Techniques for verifying the accuracy of risk measurement models," *Journal of Derivatives*, **2**, 173-84.

Read, M (2011) Quantile modeling and structural breaks in financial trade durations. Honours thesis in Econometrics, The University of Sydney (available upon request).

Smith, M., and Kohn, R. (1996). "Non-parametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, **75**, 317-343.

Tuyl, F., and Gerlach, R., and Mengersen, K. (2008). "Inference for Proportions in a 2 x 2 Contingency Table: HPD or not HPD?," *Biometrics*, **64**, 1293-1296.