Backtesting Trading Book Models

Using VaR, Expected Shortfall and Realized *p*-Values

Alexander J. McNeil[1]

[1]Heriot-Watt University, Edinburgh

Vienna
10 June 2015

Overview

Overview

The Trading Book

- Contains assets that are available to trade.
- Can be contrasted with the more traditional banking book which contains loans and other assets that are typically held to maturity and not traded.
- Trading book is supposed to contain assets that can be assigned a fair value at any point in time based on “marking to market” or “marking to model”.
- Examples: fixed income instruments; derivatives.
- The trading book is often identified with market risk whereas the banking book is largely affected by credit risk.
- The Basel rules allow banks to use internal Value-at-Risk (VaR) models to measure market risks in the trading book.
- These models are used to estimate a P&L (profit-and-loss) distribution from which risk measures like VaR (value-at-risk) and ES (expected shortfall) are calculated.
- Risk measures are used to determine regulatory capital requirements and for internal limit setting.

Trading Book Losses

- The risk factors at time t are denoted by the vector $\boldsymbol{Z}_t = (Z_{t,1}, \ldots, Z_{td,})$. These include, for example, equity prices, exchange rates, interest rates for different maturities and volatility parameters.
- The value of the trading book is given by formula of form

$$V_t = f_{[t]}(t, \boldsymbol{Z}_t) \tag{1}$$

  where $f_{[t]}$ is the portfolio mapping at t which is assumed to be known.
- The risk factors $\boldsymbol{Z}_t$ are observable at time t and hence V_t is known at t.
- Assuming positions held over the period $[t, t+1]$ and ignoring intermediate income, the trading book loss is described by

$$\begin{aligned} L_{t+1} = -(V_{t+1} - V_t) &= -\left(f_{[t]}(t+1, \boldsymbol{Z}_{t+1}) - f_{[t]}(t, \boldsymbol{Z}_t)\right) \\ &= -\left(f_{[t]}(t+1, \boldsymbol{Z}_t + \boldsymbol{X}_{t+1}) - f_{[t]}(t, \boldsymbol{Z}_t)\right) \\ &= l_{[t]}(\boldsymbol{X}_{t+1}) \end{aligned}$$

  where $\boldsymbol{X}_{t+1} = \boldsymbol{Z}_{t+1} - \boldsymbol{Z}_t$ are the risk-factor changes and $l_{[t]}$ is a function we refer to as the loss operator.

Estimating VaR and ES

- The bank (ideally) estimates the conditional loss distribution

$$F_{L_{t+1}|\mathcal{F}_t}(x) = P\left(l_{[t]}(\boldsymbol{X}_{t+1}) \leq x \mid \mathcal{F}_t\right)$$

  where $\mathcal{F}_t$ denotes the available information at time t. Typically this is the information in past risk-factor changes $\mathcal{F}_t = \sigma(\{(\boldsymbol{X}_s) : s \leq t\})$.
- Some methods used in practice (e.g. historical simulation) apply an unconditional approach, assuming stationarity of past risk-factor changes $(\boldsymbol{X}_s)_{s\leq t}$ and estimating the distribution of $l_{[t]}(\boldsymbol{X})$ under stationary distribution $F_{\boldsymbol{X}}$.
- The bank's forms an estimate $\widehat{F}_{L_{t+1}|\mathcal{F}_t}$ of the loss distribution using historical data up to time t. The estimate is intended to be particularly accurate in the tail.
- We write VaR_α^t and ES_α^t for the α-quantile and α-shortfall of the true conditional loss distribution $F_{L_{t+1}|\mathcal{F}_t}$ and we write $\widehat{\text{VaR}}_\alpha^t$ and $\widehat{\text{ES}}_\alpha^t$ for estimates of these quantities based on the model $\widehat{F}_{L_{t+1}|\mathcal{F}_t}$.
- The model may be parametric or non-parametric (based on the empirical distribution function).

VaR and ES: Reminder

Let F_L denote a generic loss df and let $0 < \alpha < 1$. Typically $\alpha \geq 0.95$.

- Value at Risk is defined to be

$$\text{VaR}_\alpha = q_\alpha(F_L) = F_L^{\leftarrow}(\alpha)\,, \tag{2}$$

  where we use the notation $q_\alpha(F_L)$ for a quantile of the distribution and $F_L^{\leftarrow}$ for the (generalized) inverse of F_L.

- Provided the integral converges, expected shortfall is defined to be

$$\text{ES}_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 q_u(F_L)du. \tag{3}$$

  If F_L is a continuous df and $L \sim F_L$ then

$$\text{ES}_\alpha = E(L \mid L > \text{VaR}_\alpha(L))\,.$$

We will assume the true underlying loss distributions are continuous.

Overview

The Backtesting Problem

- The estimates $\widehat{\text{VaR}}_\alpha^t$ and $\widehat{\text{ES}}_\alpha^t$ derived at time t for the loss operator $l_{[t]}$ and time horizon $[t, t+1]$ are compared with the realization of L_{t+1} at time $t+1$.
- This is a one-off, unrepeatable experiment because the loss operator $l_{[t]}$ and the conditional distribution $F_{\boldsymbol{X}_{t+1}|\mathcal{F}_t}$ change at each time point.
- In fact the idea of a "true distribution" $F_{L_{t+1}|\mathcal{F}_t}$ is abstract given that we only ever see one observation from this distribution. Davis (2014) refers to any hypothesis that $F_{L_{t+1}|\mathcal{F}_t}$ takes a particular specified form as being unfalsifiable and therefore meaningless.
- Nevertheless, we adhere to the idea of a true underlying model at each time point.
- Even if we can never reject a hypothesized model at a particular time point t, we can collect evidence over time that we have a tendency to use models with a particular deficiency (e.g. a tendency to underestimate VaR).

Overview

Violations and Their Properties

- The event $\{L_{t+1} > \text{VaR}_\alpha^t\}$ is a (theoretical) VaR violation or exception.
- Define the event indicator variable by $I_{t+1} = I_{\{L_{t+1} > \text{VaR}_\alpha^t\}}$.
- By definition of the quantile and continuity of $F_{L_{t+1}|\mathcal{F}_t}$ we have

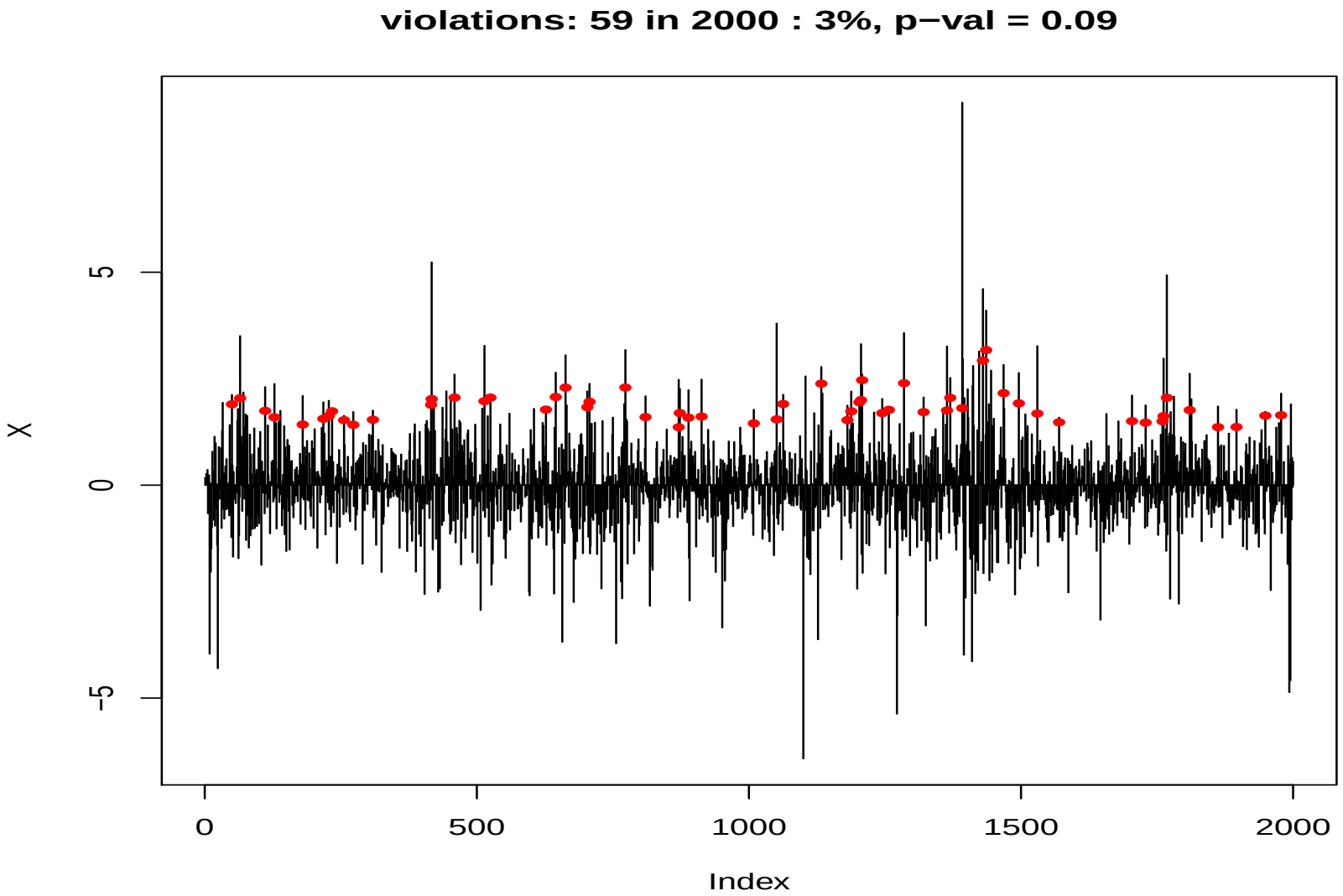
$$E(I_{t+1} \mid \mathcal{F}_t) = P(L_{t+1} > \text{VaR}_\alpha^t \mid \mathcal{F}_t) = 1 - \alpha\,. \tag{4}$$

- It may be shown that a process $(I_t)_{t\in\mathbb{Z}}$ adapted to $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ and satisfying $E(I_t \mid \mathcal{F}_{t-1}) = 1 - \alpha$ for all t is a Bernoulli trials process (iid variables).
- Implication 1: $M = \sum_{t=1}^m I_{t+1} \sim B(m, 1-\alpha)$.
- Implication 2: Let $T_0 = 0$ and define the violation times by

$$T_j = \inf\{t : T_{j-1} < t, L_{t+1} > \text{VaR}_\alpha^t\}, \quad j = 1, 2, \ldots.$$

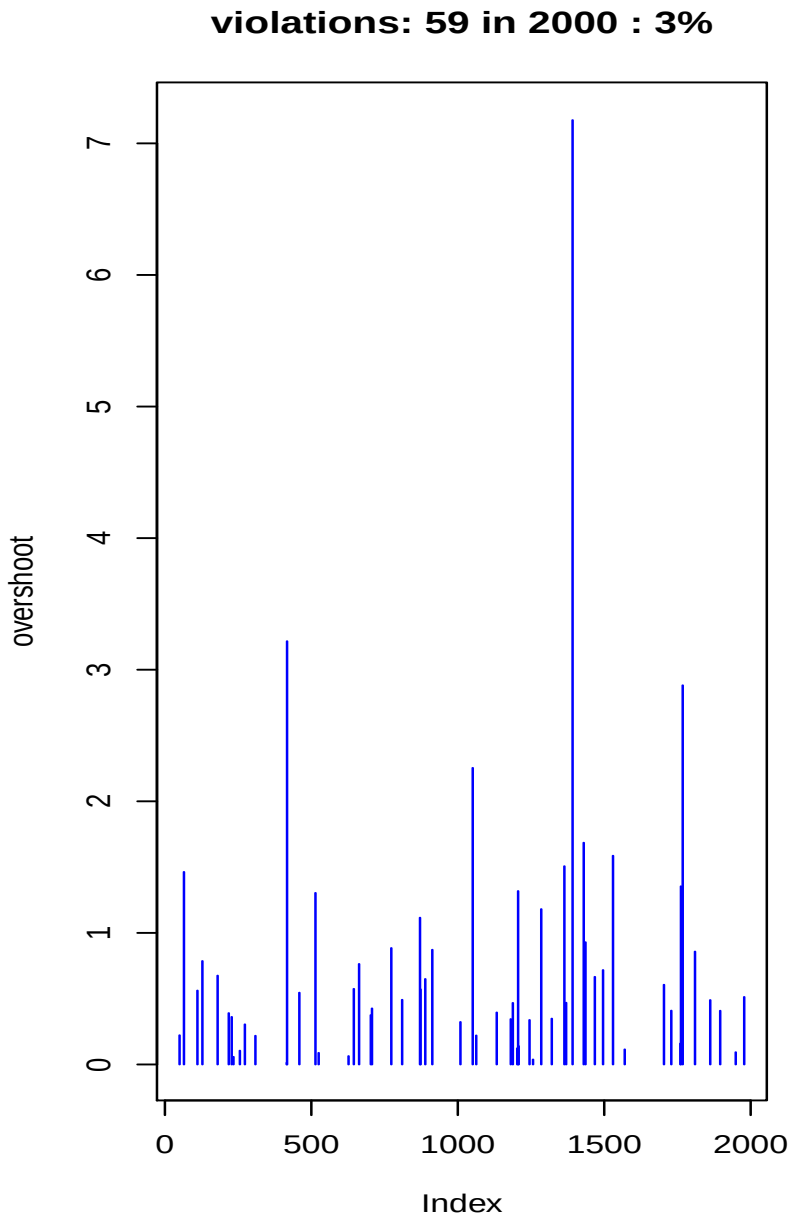
The spacings $S_j = T_j - T_{j-1}$ are independent geometric random variables with mean $1/(1-\alpha)$.

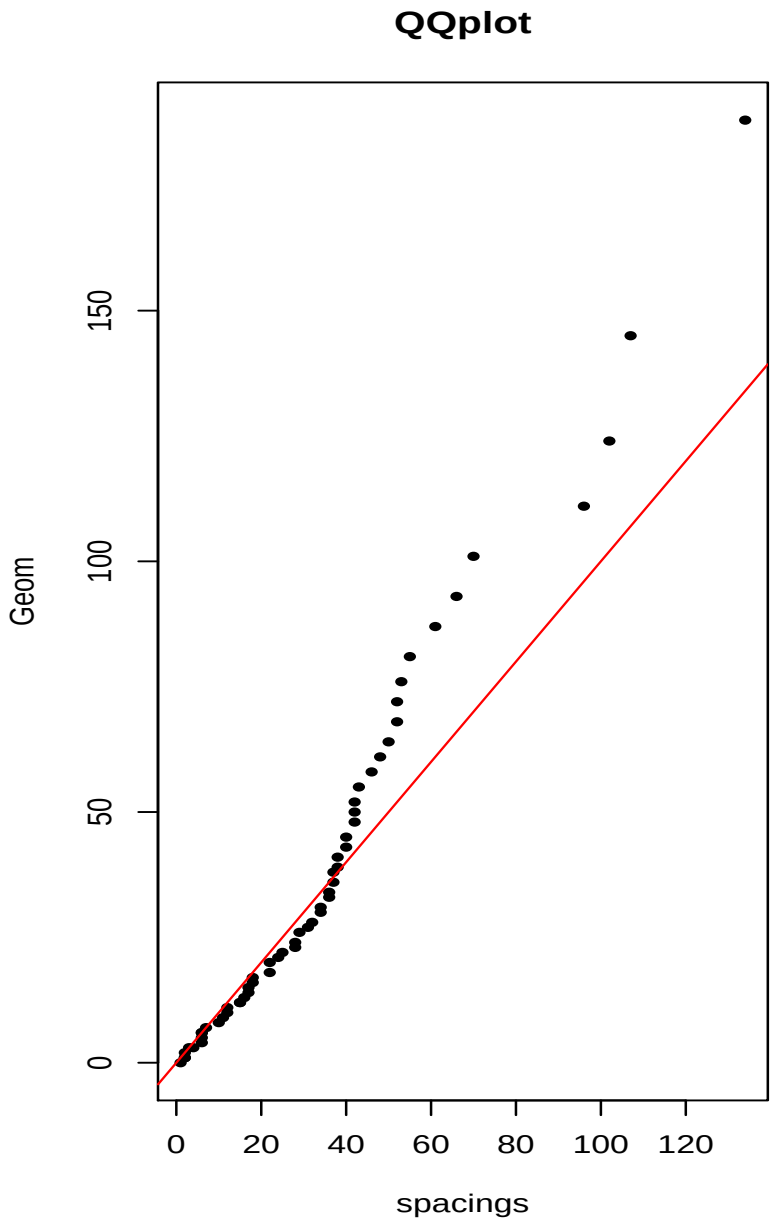
Theoretical Violations in GARCH Process



Here we consider $\text{VaR}_{0.975}$.

Point Process of Violations + Spacings





Calibration Function or Signature

- Following Christoffersen (1998) a test of the binomial behaviour of the number of violations is often referred to as a test of unconditional coverage and a test that also addresses the hypothesis of independence is a test of conditional coverage.
- The property (4) can be expressed in terms of a calibration function (Davis, 2014) (also known as a signature in elicitability literature).
- That is, we may write (4) as

$$E\left(h_\alpha(\text{VaR}_\alpha^t, L_{t+1}) \mid \mathcal{F}_t\right) = 0$$

where h_α is the calibration function given by

$$h_\alpha(q, l) = I_{\{l>q\}} - (1-\alpha). \tag{5}$$

- Remarkably $(h_\alpha(\text{VaR}_\alpha^t, L_{t+1}))$ forms a process of mean-zero iid variables regardless of underlying model.

Overview

Formulating Hypotheses

- In practice VaR_α^t is estimated at a series of time points $t = 1, \ldots, m$ and we test the null and alternative hypotheses

$$H_0 \quad : \quad E\left(h_\alpha(\widehat{\text{VaR}}_\alpha^t, L_{t+1}) \mid \mathcal{F}_t\right) = 0, \quad t = 1, \ldots, m,$$

$$H_1 \quad : \quad E\left(h_\alpha(\widehat{\text{VaR}}_\alpha^t, L_{t+1}) \mid \mathcal{F}_t\right) \geq 0, \quad t = 1, \ldots, m \quad \text{(with } > \text{ for some } t\text{)}.$$

- The null is equivalent to the hypothesis that VaR_α^t is correctly estimated at all time points and the alternative is that VaR_α^t is systematically underestimated.
- Under H_0 we have

$$P(L_{t+1} > \widehat{\text{VaR}}_\alpha^t \mid \mathcal{F}_t) = 1 - \alpha\,,$$

- Thus the violation indicator variables defined by $\widehat{I}_{t+1} = I_{\{L_{t+1} > \widehat{\text{VaR}}_\alpha^t\}}$ form a Bernoulli trials process with event probability α.

Possible Tests

- Tests are based on the realized values of $\{\widehat{I}_{t+1} : t = 1, \ldots, m\}$.
- If we define the statistic $\sum_{t=1}^{m} \widehat{I}_{t+1}$ then, under H_0, this statistic should have a binomial distribution.
- A test for binomial behaviour can be based on a likelihood ratio statistic (Christoffersen, 1998), score statistic or direct comparison with binomial.
- Christoffersen (1998) proposed a test for independence of violations against the alternative of first-order Markov behaviour; a similar test is considered in Davis (2014).
- Christoffersen and Pelletier (2004) proposed a test based on the spacings between violations. The null hypothesis of exponential spacings (constant hazard model) is tested against a Weibull alternative (in which the hazard function may be increasing or decreasing). See also Berkowitz et al. (2011).
- A regression-based approach using the CAViaR framework of Engle and Manganelli (2004) works well.

Overview

Finding a Calibration Function for Expected Shortfall

- A natural approach to backtesting expected shortfall estimates is to look for a calibration function, that is a function h such that

$$E(h(\mathrm{ES}_\alpha^t, L_{t+1}) \mid \mathcal{F}_t) = 0$$

  for a large class of models.
- However, it is not possible to find such a function (a fact that is related to the non-elicitability of expected shortfall; see Acerbi and Szekely (2014)).
- Instead, the backtests that have been proposed generally rely on calibration functions h that also reference VaR and satisfy

$$E(h(\mathrm{VaR}_\alpha^t, \mathrm{ES}_\alpha^t, L_{t+1}) \mid \mathcal{F}_t) = 0.$$

First Calibration Function

- By the definition of expected shortfall we have that

$$E\left((L_{t+1} - \mathrm{ES}_\alpha^t) I_{t+1} \mid \mathcal{F}_t\right) = 0\,. \tag{6}$$

- Using the calibration function

$$h^{(1)}(q, e, l) = \left(\frac{l - e}{e}\right) I_{\{l>q\}}$$

we define the quantity

$$K_{t+1} = h^{(1)}(\mathrm{VaR}_\alpha^t, \mathrm{ES}_\alpha^t, L_{t+1})\,. \tag{7}$$

- Expressions of this kind were studied in McNeil and Frey (2000) who used them to define violation residuals.
- The idea of analysing (7) has been further developed in Acerbi and Szekely (2014). Clearly we have that

$$E(K_{t+1}) = E\left(K_{t+1} \mid \mathcal{F}_t\right) = 0.$$

Second Calibration Function

- Acerbi and Szekely (2014) obtained an alternative calibration function by considering

$$E\left(L_{t+1} I_{t+1} \mid \mathcal{F}_t\right) - \mathrm{ES}_\alpha^t(1-\alpha) = 0\,, \tag{8}$$

  which also follows from (6).
- If we define

$$h_\alpha^{(2)}(q, e, l) = \frac{I_{\{l>q\}}}{e} - (1-\alpha)$$

  we can set

$$S_{t+1} = h_\alpha^{(2)}(\mathrm{VaR}_\alpha^t, \mathrm{ES}_\alpha^t, L_{t+1})\,, \tag{9}$$

  so that $E(S_{t+1}) = E\left(S_{t+1} \mid \mathcal{F}_t\right) = 0$.
- We use a slightly different scaling to Acerbi and Szekely (2014).
- Under our definition, S_{t+1} and K_{t+1} are related by

$$S_{t+1} = K_{t+1} + \left(I_{t+1} - (1-\alpha)\right).$$

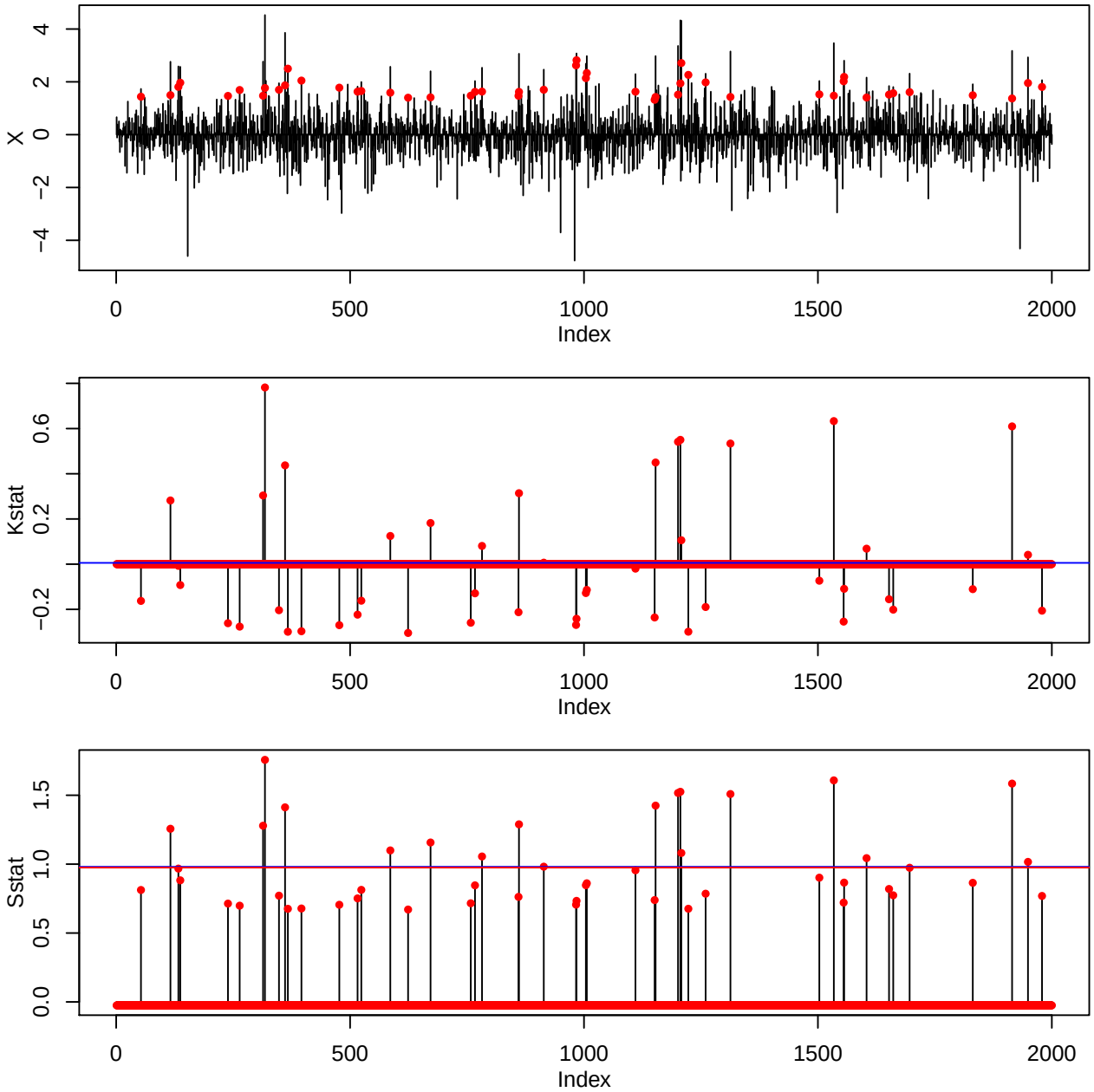
Properties of Violation Residuals

- The processes (K_t) and (S_t) are martingale difference processes ($\mathcal{F}_t$-adapted processes (Y_t) satisfying $E(Y_{t+1} \mid \mathcal{F}_t) = 0$).
- Unlike the series $(I_t - (1-\alpha))$, which is also an iid series, it is not possible to make stronger statements about (K_t) and (S_t) without making stronger assumptions about the underlying model.
- For example, suppose that losses (L_t) follow an iid innovations model of the form

$$L_t = \sigma_t Z_t, \quad \forall t, \tag{10}$$

  where $\sigma_t^2 = \text{var}(L_t \mid \mathcal{F}_{t-1})$ and (Z_t) forms a strict white noise (an iid process) with mean zero and variance one (such as a GARCH model).
- Under assumption (10) the processes (K_t) and (S_t) defined by applying the constructions (7) and (9) are processes of iid variables with mean zero.

(K_t) and (S_t) for GARCH Process ($m = 2000$)



Overview

Formulating Tests

- Now let

$$\{\widehat{K}_{t+1} = h^{(1)}(\widehat{\text{VaR}}_\alpha^t, \widehat{\text{ES}}_\alpha^t, L_{t+1}) \quad : \quad t = 1, \ldots, m\}$$
$$\{\widehat{S}_{t+1} = h_\alpha^{(2)}(\widehat{\text{VaR}}_\alpha^t, \widehat{\text{ES}}_\alpha^t, L_{t+1}) \quad : \quad t = 1, \ldots, m\}$$

  denote the violation residuals obtained when estimates of VaR_α^t and ES_α^t are inserted in the calibration functions.
- We consider the problem of testing for mean-zero behaviour in these residuals.
- Clearly we have the relationship

$$\widehat{S}_{t+1} = \widehat{K}_{t+1} + h_\alpha(\widehat{\text{VaR}}_\alpha^t, L_{t+1}) \tag{11}$$

  where h_α is the calibration function for VaR estimation.
- A test for the mean-zero behaviour of the $\widehat{S}_{t+1}$ residuals can be thought of as combining a test for the mean-zero behaviour of the $\widehat{K}_{t+1}$ residuals and a test for correct VaR estimation.

Mean-Zero Test for $(\widehat{K}_t)$ Residuals

- Hypotheses:

$$\begin{aligned} H_0 &: \widehat{F}_{L_{t+1}|\mathcal{F}_t}(x) = F_{L_{t+1}|\mathcal{F}_t}(x), \quad x \geq \mathrm{VaR}_\alpha^t, \ t = 1, \ldots, m, \\ H_1 &: E(\widehat{K}_{t+1}) \geq 0, \quad t = 1, \ldots, m \quad (\text{with} > \text{for some } t). \end{aligned}$$

- Null implies that VaR and ES are correctly estimated and $E(\widehat{K}_{t+1}) = 0$.
- Alternative can arise from different deficiencies of $\widehat{F}_{L_{t+1}|\mathcal{F}_t}$; true for example if VaR is correctly estimated but ES underestimated.
- A test based on the $(\widehat{K}_t)$ residuals could be viewed as a second-stage test after the null hypothesis of accurate VaR estimation had been tested.
- We note that

$$E(\widehat{K}_{t+1}) = 0 \iff E(\widehat{K}_{t+1} \mid L_{t+1} > \widehat{\mathrm{VaR}}_\alpha^t) = 0.$$

- It suffices to test the values of $\widehat{K}_{t+1}$ at times when violations occur for mean-zero behaviour.

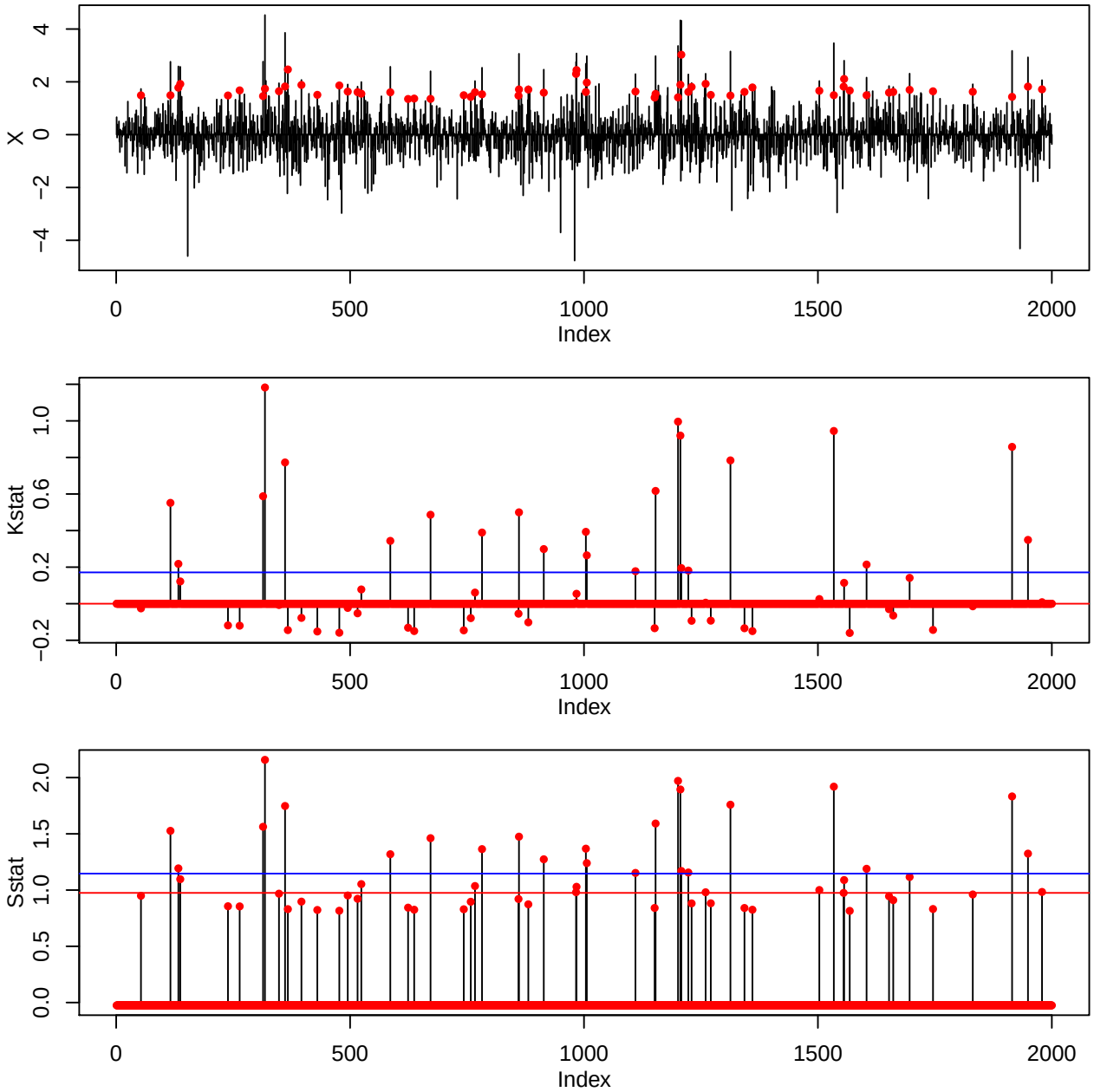
Bootstrap Test or t-Test

- McNeil and Frey (2000) suggest a bootstrap hypothesis test of H_0 against H_1 based on the non-zero residuals.
- This is an example of a one-sample bootstrap hypothesis test as described by Efron and Tibshirani (1994) (page 224).
- A standard one-sample t test could also be carried out.
- In using such tests we implicitly assume that the residuals form an identically distributed sample.
- This would be true under the null hypothesis if we also assume an iid innovations model structure as in (10).

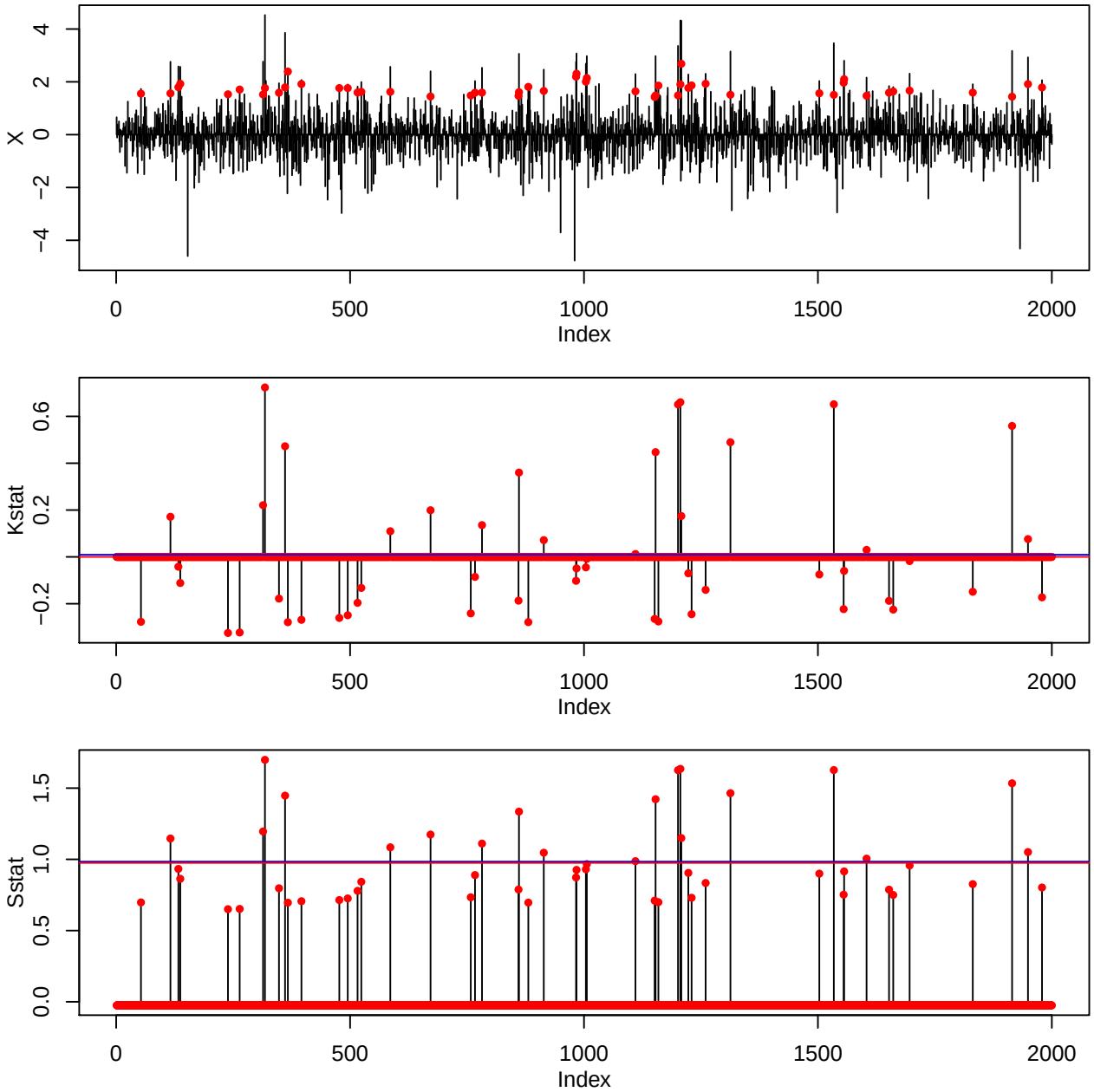
Example

- Simulation Experiment. The true data generating mechanism is a GARCH(1,1) model with Student *t* innovations with 4 degrees of freedom.
- Models are estimated using windows of 1000 past data but are only refitted every 10 steps.
- Model A. Forecaster uses an ARCH(1) model with normal innovations. This is misspecified with respect to the form of the dynamics and the distribution of the innovations.
- Model B. Forecaster uses a GARCH(1,1) model with normal innovations. This is misspecified with respect to the distribution of the innovations.
- Model C. Forecaster uses a GARCH(1,1) model with Student *t* innovations. He has identified correct dynamics and distribution but still has to estimate the parameters of model.
- The aim is to estimate the 97.5% VaR and expected shortfall of $F_{X_{t+1}|\mathcal{F}_t}$.
- Binomial test p-values for A, B and C are 0.21, 0.07, 0.35.
- Shortfall t-test p-values for A, B and C are 0.00, 0.00, 0.41.

Residuals Model B



Residuals Model C



Overview

Acerbi-Szekely Test

- Acerbi and Szekely (2014) suggest the use of a Monte Carlo hypothesis test; see Davison and Hinkley (1997) (page 140).
- This test may be applied to either set of residuals and we describe its application to $\{\widehat{S}_{t+1} : t = 1, \ldots, m\}$.
- We consider the hypotheses

$$\begin{aligned} H_0 &: \widehat{F}_{L_{t+1}|\mathcal{F}_t}(x) = F_{L_{t+1}|\mathcal{F}_t}(x), \quad x \geq \text{VaR}_\alpha^t, \ t = 1, \ldots, m, \\ H_1 &: E(\widehat{S}_{t+1}) \geq 0, \quad t = 1, \ldots, m \quad (\text{with} > \text{for some } t). \end{aligned} \tag{12}$$

- The observed value for the test statistic is

$$S_0 = m^{-1} \sum_{t=1}^{m} \widehat{S}_{t+1} = m^{-1} \sum_{t=1}^{m} h_\alpha^{(2)}(\widehat{\text{VaR}}_\alpha^t, \widehat{\text{ES}}_\alpha^t, L_{t+1})$$

- We generate a random sample from the distribution of S_0 under the null hypothesis and compare with S_0

Acerbi-Szekely Test Procedure

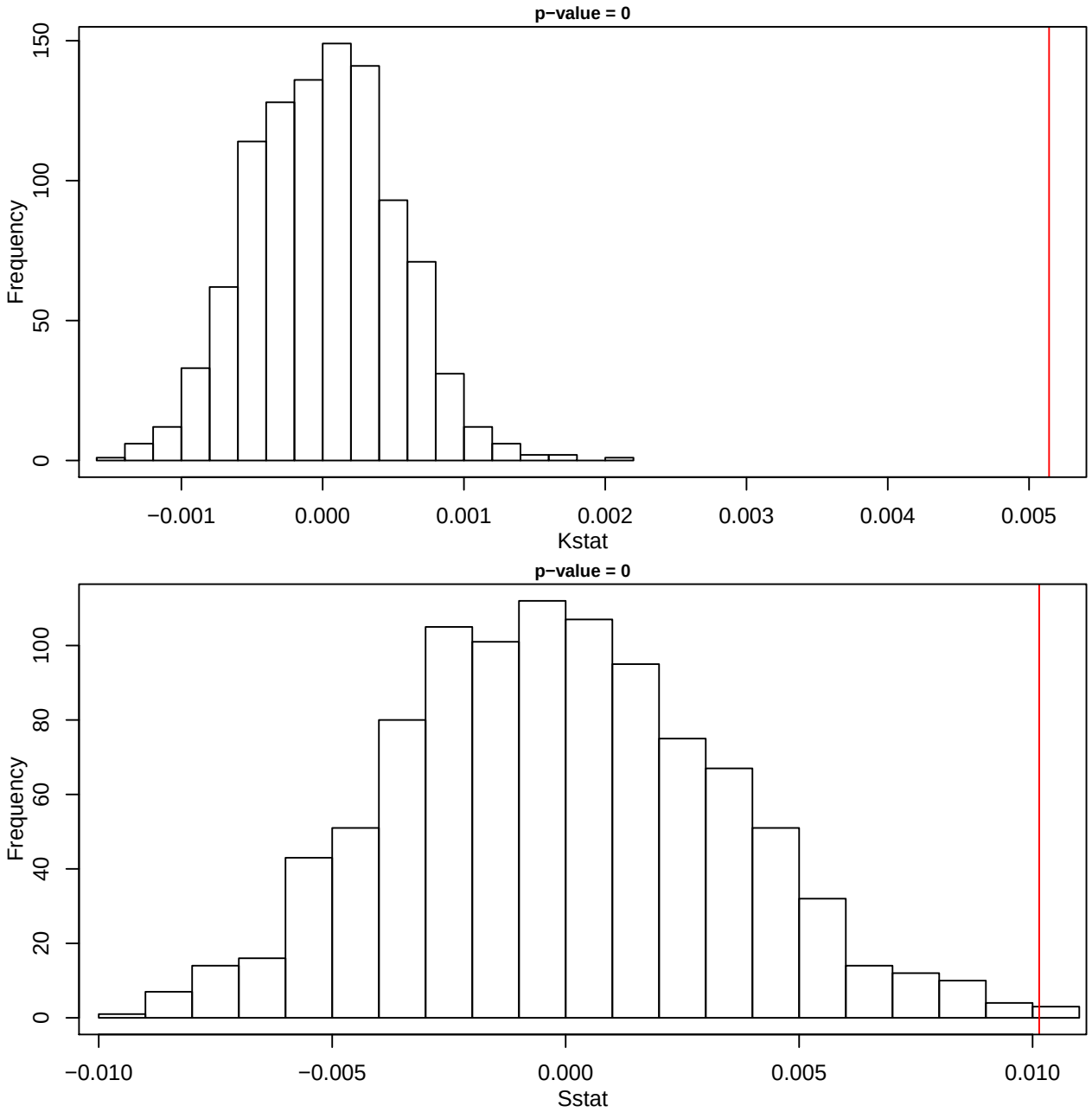
1. We generate $L_{t+1}^{(j)}$ from $\widehat{F}_{L_{t+1}|\mathcal{F}_t}$ for $t = 1, \ldots, m$ and $j = 1, \ldots, n$. Since only the tail of the model is specified under H_0 and the test statistic does not depend on the exact values of $L_{t+1}^{(j)}$ when $L_{t+1}^{(j)} \leq \widehat{\text{VaR}}_\alpha^t$, it suffices to generate any value $k \leq \widehat{\text{VaR}}_\alpha^t$ with probability α and a value from the conditional distribution $\widehat{F}_{L_{t+1}|L_{t+1}>\widehat{\text{VaR}}_\alpha^t,\mathcal{F}_t}$ with probability $(1-\alpha)$.
2. For each Monte Carlo sample indexed by j we compute

$$S^{(j)} = m^{-1} \sum_{t=1}^{m} h_\alpha^{(2)}(\widehat{\text{VaR}}_\alpha^t, \widehat{\text{ES}}_\alpha^t, L_{t+1}^{(j)}).$$

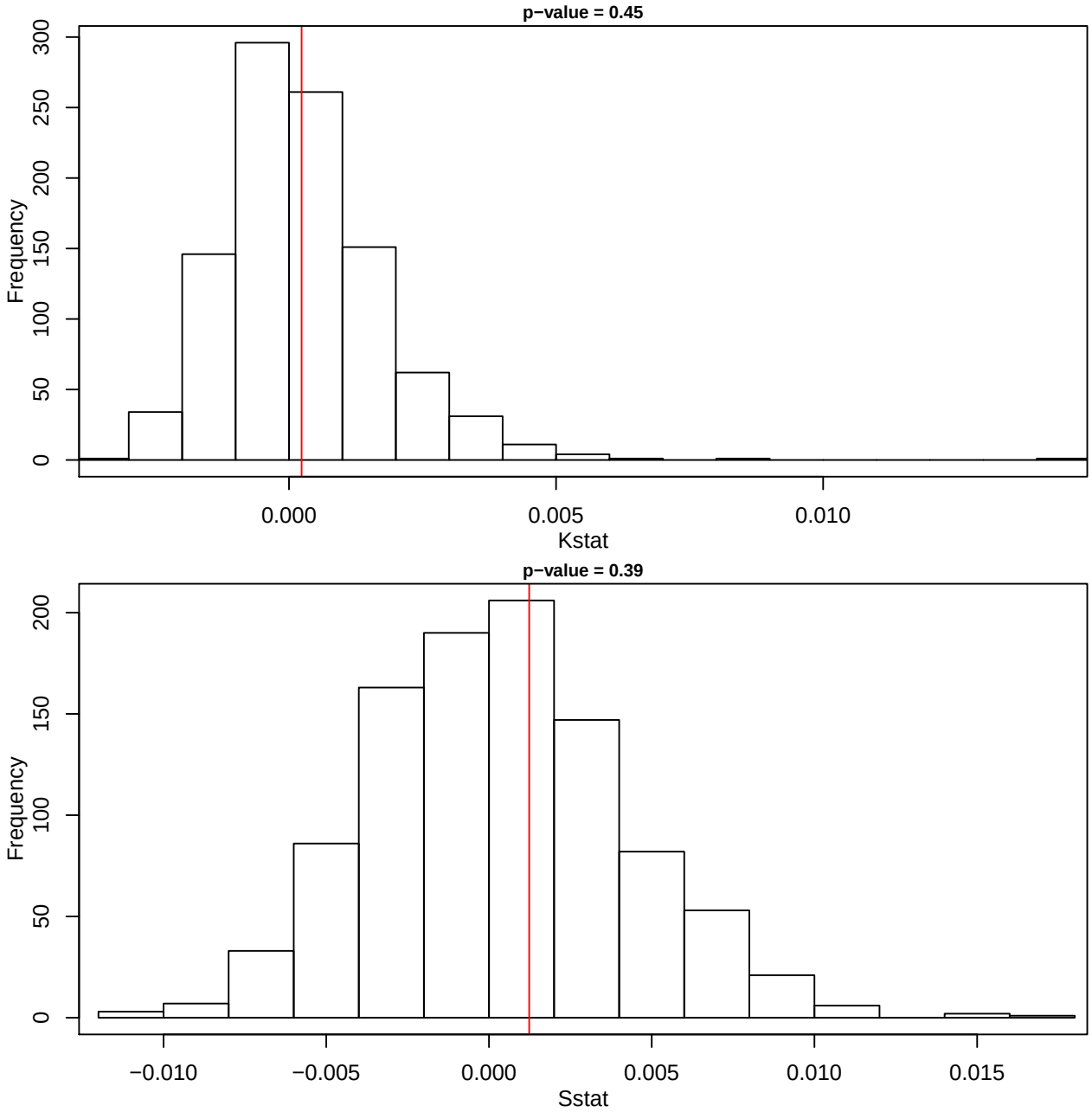
3. Estimate p-value by fraction of the values $S_0, S^{(1)}, \ldots, S^{(n)}$ greater than or equal to S_0.

This test has the advantage that we do not have to assume the residuals are identically distributed.
It has the disadvantage that we have to record details of the tail models used at each time point in order to generate Monte Carlo samples.

Results Model B



Results Model C



Overview

Elicitability and Scoring Functions

- The elicitability concept has been introduced into the backtesting literature by Gneiting (2011); see also important papers by Bellini and Bignozzi (2013) and Ziegel (2015).
- A key concept is that of a scoring function $S(y, l)$ which measures the discrepancy between a forecast y and a realized loss l.
- Forecasts are made by applying real-valued statistical functionals T (such as mean, median or other quantile) to the distribution of the loss F_L to obtain the forecast $y = T(F_L)$.
- Suppose that for some class of loss distribution functions a real-valued statistical functional T satisfies

$$T(F_L) = \arg\min_{y \in \mathbb{R}} \int_{\mathbb{R}} S(y, l) \mathrm{d}F_L(l) = \arg\min_{y \in \mathbb{R}} E(S(y, L)) \tag{13}$$

  for a scoring function S and any loss distribution F_L in that class.
- Suppose moreover that $T(F_L)$ is the unique minimizing value.

Elicitability and Calibration Functions

- The scoring function S is said to be strictly consistent for T.
- The functional $T(F_L)$ is said to be elicitable.
- Note that (13) implies that

$$\frac{\mathrm{d}}{\mathrm{d}y} E(S(y, L))\Big|_{y=T(F_L)} = \int_{\mathbb{R}} \frac{\mathrm{d}}{\mathrm{d}y} S(y, l)\mathrm{d}F_L(l)\Big|_{y=T(F_L)} = E(h(T(F_L), L)) = 0$$

  where h is the derivative of the scoring function.
- Thus elicitability theory also indicates how we may derive calibration functions for hypothesis tests involving $T(F_L)$.

Elicitability: Examples

- The VaR risk measure corresponds to $T(F_L) = F_L^{\leftarrow}(\alpha)$. For any $0 < \alpha < 1$ this functional is elicitable for strictly increasing distribution functions. The scoring function

$$S_\alpha^q(y, l) = |1_{\{l \leq y\}} - \alpha||l - y| \tag{14}$$

  is strictly consistent for T.
- If we take the negative of the derivative of this function with respect to y we get the calibration function $h_\alpha(y, l)$ in (5).
- The α-expectile of L is defined to be the risk measure that minimizes $E\left(S_\alpha^e(y, L)\right)$ where the scoring function is

$$S_\alpha^e(y, l) = |1_{\{l \leq y\}} - \alpha|(l - y)^2. \tag{15}$$

  This risk measure is elicitable by definition.
- Bellini and Bignozzi (2013) and Ziegel (2015) show that a risk measure is coherent and elicitable if and only if it is the α-expectile risk measure for $\alpha \geq 0.5$; see also Weber (2006). Expected shortfall is not elicitable.

Elicitability in Backtesting Context

- VaR_α^t minimizes

$$E\left(S_\alpha^q(\text{VaR}_\alpha^t, L_{t+1}) \mid \mathcal{F}_t\right)$$

for the scoring function in (14). We refer to $S_\alpha^q(\text{VaR}_\alpha^t, L_{t+1})$ as a (theoretical) VaR score.

- The (theoretical) VaR scores for the realization of the GARCH process can be calculated.
- For the GARCH process it may be shown that

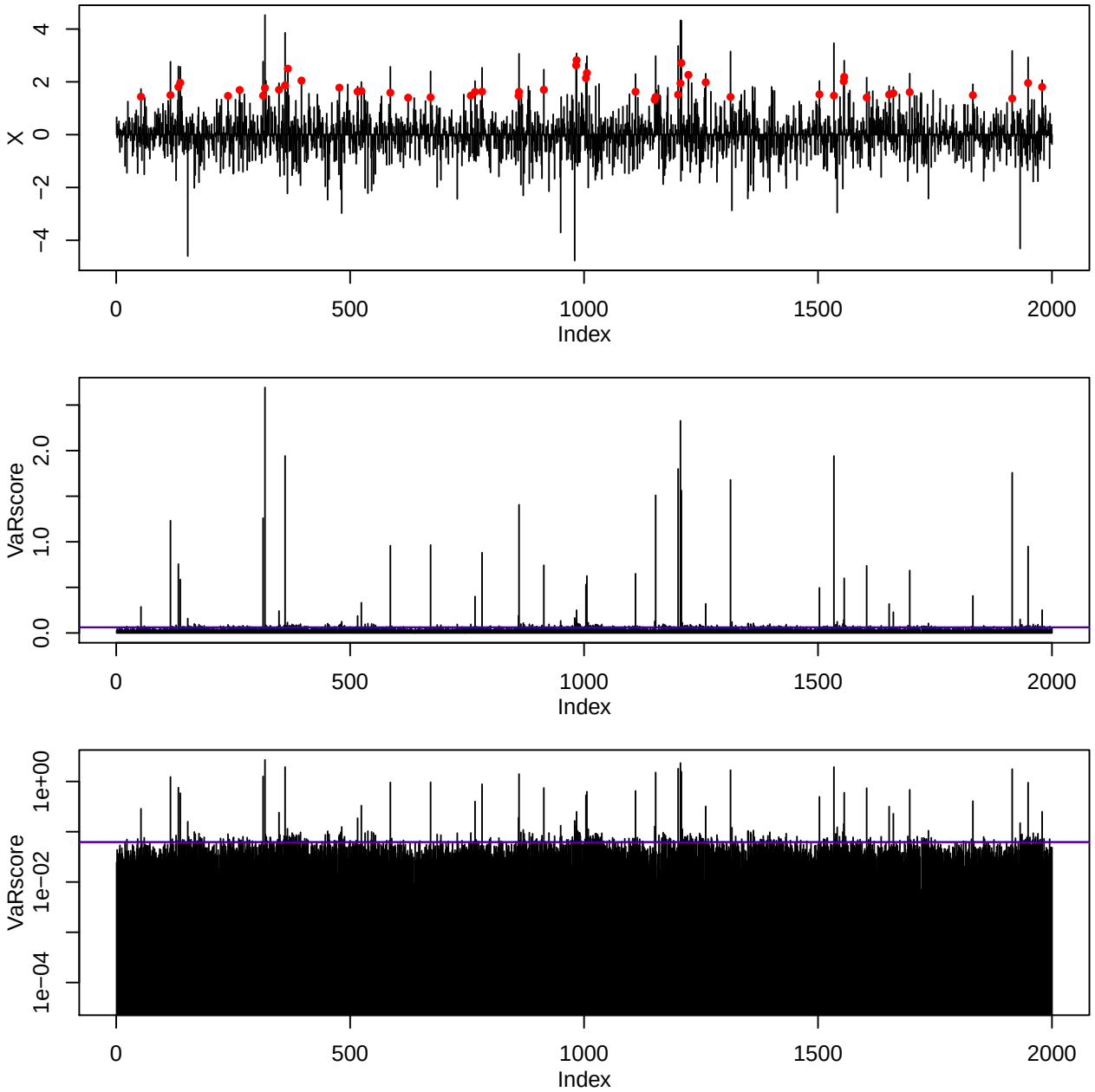
$$S_\alpha^q(\text{VaR}_\alpha^t, L_{t+1}) = \sigma_{t+1} S_\alpha^q(q_\alpha(Z), Z_{t+1})$$

where $q_\alpha(Z)$ denotes the α-quantile of the innovation distribution.

- Since the theoretical VaR scores form a stationary and ergodic process

$$\lim_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} S_\alpha^q(\text{VaR}_\alpha^t, L_{t+1}) = E(\sigma) E\left(S_\alpha^q(q_\alpha(Z), Z)\right).$$

VaR Scores for GARCH



Overview

Model Comparison

- Assume VaR_α^t is replaced by an estimate at each time point and consider the VaR scores $\{S_\alpha^q(\widehat{\text{VaR}}_\alpha^t, L_{t+1}) : t = 1, \ldots, m\}$
- These can be used to address questions of relative and absolute model performance.
- The statistic

$$Q_0 = \frac{1}{m} \sum_{t=1}^{m} S_\alpha^q(\widehat{\text{VaR}}_\alpha^t, L_{t+1})$$

  can be used as a measure of relative model performance.
- If two models A and B deliver VaR estimates $\{\widehat{\text{VaR}}_\alpha^{t\,A}, t = 1, \ldots, m\}$ and $\{\widehat{\text{VaR}}_\alpha^{t\,B}, t = 1, \ldots, m\}$ with corresponding average scores Q_0^A and Q_0^B, then we expect the better model to give estimates closer to the true VaR numbers and thus a value of Q_0 that is lower.
- Of course, the power to discriminate between good models and inferior models will depend on the length of the backtest.

Overview

Model Validation

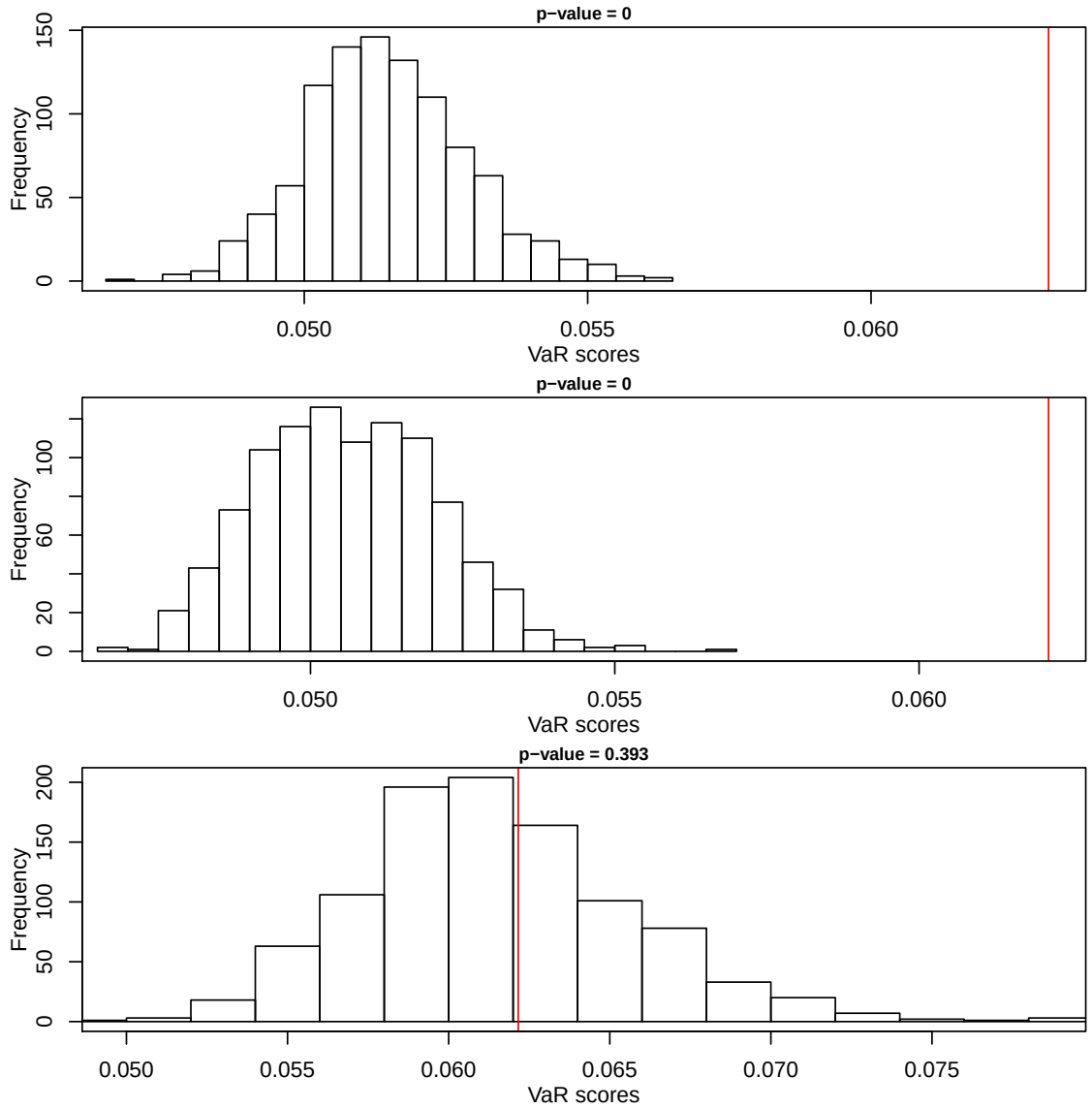
- We can also consider the question of whether a score indicates that any particular model is good enough.
- One approach to this problem is to use the score as the basis of a goodness-of-fit test. The hypotheses could be formulated as

$$\begin{aligned} H_0 &: \widehat{F}_{L_{t+1}|\mathcal{F}_t} = F_{L_{t+1}|\mathcal{F}_t}, \quad t = 1, \ldots, m, \\ H_1 &: \widehat{F}_{L_{t+1}|\mathcal{F}_t} \neq F_{L_{t+1}|\mathcal{F}_t}, \quad \text{for at least some } t. \end{aligned}$$

- Note that the model is fully specified under the null hypothesis in contrast to the test of tail fit set out in (12). This framework allows us to carry out the following Monte Carlo test.

1. We generate $L_{t+1}^{(j)}$ under H_0. That is we generate $L_{t+1}^{(j)}$ from $\widehat{F}_{L_{t+1}|\mathcal{F}_t}$ for $t = 1, \ldots, m$ and $j = 1, \ldots, n$.
2. For each Monte Carlo sample we compute $Q^{(j)} = m^{-1} \sum_{t=1}^{m} S_\alpha^q(\widehat{\text{VaR}}_\alpha^t, L_{t+1}^{(j)})$.
3. We estimate p-value by fraction of the values $Q_0, Q^{(1)}, \ldots, Q^{(n)}$ that are greater or equal to Q_0.

Monte Carlo Goodness-of-Fit Using VaR Scores



Models A, B and C.

Overview

Realized p-Values

- We briefly consider an alternative to backtests based on expected shortfall.
- Let $U_{t+1} = F_{L_{t+1}|\mathcal{F}_t}(L_{t+1})$ for $t = 0, 1, 2, \ldots$. Under continuity assumptions, the process $(U_t)_{t\in\mathbb{N}}$ is a process of iid standard uniform variables.
- Denoting the estimated model at time t by $\widehat{F}_{L_{t+1}|\mathcal{F}_t}$ as usual, we define realized p-values by $\widehat{U}_{t+1} = \widehat{F}_{L_{t+1}|\mathcal{F}_t}(L_{t+1})$ for $t = 0, 1, 2, \ldots$.
- Realized p-values effectively contain information about VaR violations at any level α:

$$\widehat{U}_{t+1} > \alpha \iff L_{t+1} > \widehat{F}^{\leftarrow}_{L_{t+1}|\mathcal{F}_t}(\alpha)$$

  if $\widehat{F}_{L_{t+1}|\mathcal{F}_t}$ is strictly increasing and continuous.
- It is possible to transform uniform variables to any scale. For example, if we define $\widehat{Z}_{t+1} = \Phi^{-1}(\widehat{U}_{t+1})$, where Φ is the standard normal df, then we would expect that the $(\widehat{Z}_t)$ variables are iid standard normal. Berkowitz (2001) has proposed a test based on this fact.

Berkowitz Test

- The realized *p*-values can be truncated by defining

$$\widehat{U}^*_{t+1} = \min\left(\max\left(\widehat{U}_{t+1}, \alpha_1\right), \alpha_2\right) \quad 0 \leq \alpha_1 < \alpha_2 \leq 1.$$

- Applying the probit transformation we obtain truncated *z* values:

$$\widehat{Z}^*_{t+1} = \Phi^{-1}(\widehat{U}^*_{t+1}),\ t = 0, 1, 2, \ldots.$$

- Let $TN(\mu, \sigma^2, k_1, k_2)$ denote a normal distribution truncated to $[k_1, k_2]$.
- Under the null hypothesis of correct estimation of the loss distribution, the truncated *z*-values are iid realizations from a $TN(0, 1, \Phi^{-1}(\alpha_1), \Phi^{-1}(\alpha_2))$ distribution.
- Berkowitz applies one-sided truncation and uses a likelihood ratio test to test the null hypothesis against the alternative that the truncated *z* values have an unconstrained $TN(\mu, \sigma^2, \Phi^{-1}(\alpha_1), \infty)$ distribution.
- This can be extended to a joint test of uniformity in the tail and independence by making μ (and possibly σ) time dependent.

Overview

Conclusions I

- Value-at-Risk has special properties that make it particularly natural to backtest. Namely, the violation process forms a Bernoulli trials process under any reasonable model for the losses.
- The lack of a natural calibration function for expected shortfall, which is a consequence of the lack of elicitability, means that expected shortfall can not be backtested in isolation.
- However, it is feasible to develop joint backtests of ES and VaR.
- These can detect deficiencies of tail models that are not detected by backtesting VaR at a single level.
- The simplest tests based on expected shortfall (bootstrap test and t-test) require some additional assumptions concerning data generating mechanism.
- The Monte Carlo test of Acerbi-Szekely makes no strong assumptions but requires extensive storage of data.
- We should be aware that ES estimation procedures lack robustness.
- Tests of realized *p*-values may be an interesting alternative.

Conclusions About Use of Elicitability Theory

- Average VaR scores can be used as comparative measures to identify superior models.
- The average VaR score can also be used as the basis of a Monte Carlo goodness-of-fit test.
- Joint tests based on VaR scores at different confidence levels could be an alternative to joint tests of VaR and ES.
- The VaR score does have an attractive feature not shared by most other metrics.
- If a forecaster genuinely wanted to minimize a VaR score, he would be impelled to do the best possible job of estimating conditional quantiles of the loss distribution. It would be the optimal way to act.
- This suggests imposing financial penalties or fees on banks that are proportional to the scoring function!
- This relates to ideas of Osband (1985) about eliciting truth-telling; see also Osband and Reichelstein (1985).

For Further Reading

Acerbi, C. and Szekely, B. (2014). Back-testing expected shortfall. *Risk*, pages 1–6.

Bellini, F. and Bignozzi, V. (2013). Elicitable risk measures. Working paper, available at SSRN: http://ssrn.com/abstract=2334746.

Berkowitz, J. (2001). Testing the accuracy of density forecasts, applications to risk management. *Journal of Business & Economic Statistics*, 19(4):465–474.

Berkowitz, J., Christoffersen, P., and Pelletier, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science*, 57(12):2213–2227.

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4).

Christoffersen, P. F. and Pelletier, D. (2004). Backtesting value-at-risk: a duration-based approach. *Journal of Financial Econometrics*, 2(1):84–108.

Davis, M. H. A. (2014). Consistency of risk measure estimates. Preprint, available at arXiv:1410.4382v1.

For Further Reading (cont.)

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Engle, R. and Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

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

Osband, K. and Reichelstein, S. (1985). Information-eliciting compensation schemes. *Journal of Public Economics*, 27:107–115.

Osband, K. H. (1985). *Providing Incentives for Better Cost Forecasting*. PhD thesis, University of California, Berkeley.

For Further Reading (cont.)

Weber, S. (2006). Distribution-invariant risk measures, information, and dynamic consistency. *Mathematical Finance*, 16(2):419–441.

Ziegel, J. F. (2015). Coherence and elicitability. Mathematical Finance, doi: 10.1111/mafi.12080.