An implicit backtest for ES via a simple multinomial approach

Marie KRATZ

CREAR ESSEC Research Center on Risk

Joint work with **Yen H. LOK** (Heriot Watt Univ., Edinburgh)
& **Alexander McNEIL** (Univ. of York, UK)

Working Group on Risk
March 8, 2017

- The choice of risk measure has much impact in terms of risk management and model validation.
- Various usages of risk measures
  - The main usage of risk measures is to compute, from the probability distribution of the firm's value, the Risk Adjusted Capital in its different forms :
    - In **insurance**
      1. Solvency Capital Requirements of Solvency II : VaR (99.5% yearly)
      2. Target capital for the Swiss Solvency Test : ES (99% yearly)
    - In **banks**
      1. Basel II : VaR (99% daily)
      2. In the future Basel III : ES (97.5% daily for market risk)
  - Heart of a risk/reward strategy :
    1. to measure the diversification benefit of a risk portfolio
    2. to allow capital allocation among the various risks of the portfolio (very important role of the risk measure to optimize companies value)

- What are the main properties we should expect in practice from a "good" risk measure ?
  1. the **subadditivity** and comonotonic additivity, to measure the diversification benefit
  2. good estimates and possibility of **backtesting**

- Popular / regulatory risk measures :
  - ↪ Value-at-Risk (VaR_α) = quantile $q(\alpha)$;
  - ↪ Expected Shortfall (ES) =Tail VaR (TVaR) :

$$ES_\alpha(L) = \frac{1}{1-\alpha}\int_\alpha^1 q_\beta(L)\,d\beta \underset{F_L \text{ cont}}{=} \mathrm{E}[L \,|\, L \geq q_\alpha(L)]$$

Backtesting

1 - VaR

(a) Optimal point forecast

VaR is *elicited* by the weighted absolute error scoring function

$$s(x,y) = (\mathbf{1}_{\{x \geq y\}} - \alpha)(x-y),\ 0 < \alpha < 1 \text{ fixed}$$

(Thomson (79), Saerens (00), or Gneiting (11) for details)
$\Rightarrow$ VaR : optimal **point** forecast

$\hookrightarrow$ this allows for the *comparison of different forecast methods*.

However, in practice, we have to compare VaR predictions by a single method with observed values, in order to assess the quality of the predictions.

(b) A popular procedure : a binomial test on the proportion of violations

- Assuming a continuous loss distribution, $\mathbb{P}[L > VaR_\alpha(L)] = 1-\alpha$ $\Rightarrow$ the probability of a violation of VaR is $1-\alpha$

- We define the violation process of VaR as

$$I_t(\alpha) = \mathbf{1}_{\left\{L(t) > VaR_\alpha(L(t))\right\}}.$$

VaR forecasts are valid iff the violation process $I_t(\alpha)$ satisfies the two conditions (Christoffersen, 03) :

(**i**) $\mathbb{E}[I_t(\alpha)] = 1-\alpha$ (**ii**) $I_t(\alpha)$ and $I_s(\alpha)$ are independent for $s \neq t$

- Under (i) & (ii), $I_t(\alpha)$'s are iid $\mathcal{B}(1-\alpha)$ $\Rightarrow$ $\sum_{t=1}^{n} I_t(\alpha) \overset{d}{\sim} \mathcal{B}(n, 1-\alpha)$

In practice, it means :

- to estimate the violation process by replacing VaR by its estimates
- check that this process behaves like iid Bernoulli random variables with violation (success) probability $p_0 \simeq 1-\alpha$
- Test on the proportion p of VaR violations,

  estimated by $\frac{1}{n}\sum_{t=1}^{n} I_t(\alpha)$:

$$\text{H0} : p = p_0 = 1-\alpha \quad \text{against} \quad \text{H1} : p > p_0$$

If the proportion of VaR violations is not significantly different from $1-\alpha$, then the estimation/prediction method is reasonable.

Note :

- Convenient procedure because it can be performed **straightforwardly** within the algorithms estimating the VaR
- Condition (ii) might be violated in practice $\Rightarrow$ various tests on the independence assumption have been proposed in the literature, as e.g. one developed by Christoffersen and Pelletier (04), based on the *duration of days between the violations of the VaR thresholds*.

2 - ES

(a) Backtesting **distribution** forecasts

Testing the distribution forecasts could be helpful, in particular for tail-based risk measures like ES.

Ex : method for the out-of-sample validation of distribution forecasts, based on the Lévy-Rosenblatt transform, named also Probability Integral Transform (PIT).

See Diebold et al. ; based on the Rosenblatt result that $F(X) \overset{d}{=} \mathcal{U}(0,1)$; they observed that if a sequence of distribution forecasts coincides with the sequence of unknown conditional laws that have generated the observations, then the sequence of PIT are iid $\mathcal{U}(0,1)$.

Nevertheless, there were still some gaps to fill up before a full implementation and use in practice. Various issues left open studied by Blum (PhD thesis, 04), in part. in situations with overlapping forecast intervals and multiple forecast horizons.

(b) A component-wise optimal forecast for ES

ES : example of a risk measure whose *conditional elicitability* (see Emmer et al.) provides the possibility to forecast it in two steps.

1. We forecast the quantile (VaR_α) as

$$\hat{q}_\alpha(L) = \arg\min_x E_P[s(x, L)]$$

with $s(x, y) = (\mathbf{1}_{\{x \geq y\}} - \alpha)(x - y)$ strictly consistent scoring function

2. Fixing this value $\hat{q}_\alpha$, $E[L|L \geq \hat{q}_\alpha]$ is just an expected value. Thus we can use strictly consistent scoring function to forecast

$$\text{ES}_\alpha(L) \approx E[L|L \geq \hat{q}_\alpha].$$

If L is L^2, the score function can be chosen as the squared error :

$$\widehat{\text{ES}}_\alpha(L) \approx \arg\min_x E_{\tilde{P}}[(x - L)^2] \quad \text{where } \tilde{P}(A) = P(A|L \geq \hat{q}_\alpha).$$

(c) An implicit backtest for ES : a simple multinomial test

▷ Idea came from the following (Emmer et al.) :

$$\begin{aligned} \mathrm{ES}_\alpha(L) &= \frac{1}{1-\alpha}\int_\alpha^1 q_u(L)\,du \\ &\approx \frac{1}{4}\left[q_\alpha(L) + q_{0.75\,\alpha+0.25}(L) + q_{0.5\,\alpha+0.5}(L) + q_{0.25\,\alpha+0.75}(L)\right]. \end{aligned}$$

where $q_\alpha(L) = VaR_\alpha(L)$. Hence, if the four $q_{a\alpha+b}(L)$ are successfully backtested, then also the estimate of $\mathrm{ES}_\alpha(L)$ might be considered reliable.

▷ We can then build a backtest based on that intuitive idea of backtesting ES via **simultaneously** backtesting multiple VaR estimates evaluated with the same method as the one used to compute the ES estimate.

Note : the Basel Committee on banking Supervision suggests a variant of this ES-backtesting approach based on testing level violations for two quantiles at 97.5% and 99% level (Jan. 2016).

Building an implicit backtest for ES

Main questions :

- Does a multinomial test work better than a binomial one for model validation ?
- Which particular form of the multinomial test should we use in which situation ?
- What is the 'optimal' number of quantiles that should be used for such a test to perform well ?

To answer these questions, we build a multi-steps experiment on simulated data :

- ▷ Static view : we test distributional forms (typical for the trading book) to see if the multinomial test distinguishes well between them, in particular between their tails
- ▷ Dynamic view : looking at a time series setup in which the forecaster may misspecify both the conditional distribution of the returns and the form of the dynamics, in different ways.

A multinomial test

Testing set-up

- We have a series of ex-ante predictive models $\{F_t, t = 1, \ldots, n\}$ and a series of ex-post losses $\{L_t, t = 1, \ldots, n\}$.
- At each time t, the model F_t is used to produce estimates (or forecasts) of $VaR_{\alpha,t}$ and $ES_{\alpha,t}$ at various probability levels α.
- The VaR estimates are compared with L_t to assess the adequacy of the models in describing the losses, with particular emphasis on the most extreme losses.

We generalize the idea of Emmer et al. by considering VaR probability levels $\alpha_1, \ldots, \alpha_N$ defined, for some starting level α, by

$$\alpha_j = \alpha + \frac{j-1}{N}(1-\alpha), \quad j = 1, \ldots, N.$$

We set, for $N > 1$, $\alpha = 0.975$ (level used for ES calculation, and lowest of the two levels used for backtesting under the Basel rules for banks) and for $N = 1$, $\alpha = 0.99$ (usual level for binomial tests of VaR exceptions). We also set $\alpha_0 = 0$ and $\alpha_{N+1} = 1$.

Testing simultaneously N VaR's (with $N > 1$) leads to a multinomial distribution ; we can set the null hypothesis of the multinomial test as

$$(\text{H0}) : \; p_j := \mathbb{E}[1_{(L_t > VaR_{j,t})}] (= \mathbb{P}[L_t > VaR_{j,t}]) \;\; = \;\; p_{j,0} := 1 - \alpha_j, \quad \forall\, j = 1, \cdots, N$$

Assuming the n observations come from a loss variable L with continuous distribution F, introduce the observed cell counts between quantile levels $q_\alpha = F^{\leftarrow}(\alpha)$ as $O_j = \sum_{t=1}^{n} I_{(q_{j-1} < L_t \leq q_j)}$, for $j = 1, ..., N+1$.

Then $(O_1, \ldots, O_{N+1})$ follows a mutinomial distribution :

$$(O_1, \ldots, O_{N+1}) \sim \text{MN}(\beta_1 - \beta_0, \ldots, \beta_{N+1} - \beta_N)$$

for parameters $\beta_1 < \cdots < \beta_N$ with $\beta_0 = 0$ and $\beta_{N+1} = 1$.
Hence the test can be rewriten as

$$\left| \begin{array}{lll} \text{H0} : & \beta_j & = \alpha_j \quad \text{for } j = 1, \ldots, N \\ \text{H1} : & \beta_j & \neq \alpha_j \quad \text{for at least one } j \in \{1, \ldots, N\}. \end{array} \right.$$

To judge the relevance of the test, compute :
its size $\gamma = \mathbb{P}(\text{reject H0}) | \text{H0 true}]$ (type I error)
and its power $1 - \beta = 1 - \mathbb{P}[(\text{accept H0}) | \text{H0 wrong}]$ (1- type II error).

- Checking the size of the multinomial test : straightforward, by simulating data from a multinomial distribution under the null hypothesis (H0). This can be done by simulating data from any distribution (such as normal) and counting the observations between the true values of the α_j-quantiles, or simulating from the multinomial distribution directly.
- To calculate the power : we have to simulate data from multinomial models under the alternative hyp. (H1). Here we chose to simulate from models coming from a distribution G, having heavy tails and possibly skewness, with $G \neq F$ (true distr.), where the parameters are given by

$$\beta_j = F\left(G^{\leftarrow}(\alpha_j)\right), \quad \text{with } \beta_j \neq \alpha_j .$$

Example :
F= true distribution of L_t, so that the true quantiles $= F^{\leftarrow}(\alpha_j)$. However a modeller chooses the wrong distribution G and makes estimates $G^{\leftarrow}(\alpha_j)$ of the quantiles. The probabilities associated with these quantile estimates are s.t. $\beta_j = F(G^{\leftarrow}(\alpha_j)) \neq \alpha_j$.

This test is closely linked to the PIT method for the backtest of a probability distribution forecast. Checking that, $\forall j$, $\beta_j = F(G^{\leftarrow}(\alpha_j)) = \alpha_j$, comes back to check that $G(X_j)$ is uniformly distributed, with $X_j \sim F$, with known realizations x_j (for PIT method, we would do it for all the known realizations). Kind of PIT test.

Various test statistics can be used to describe the event (reject of H0) (see e.g. Cai and Krishnamoorthy for five possible tests for testing the multinomial proportions). Here we use, for comparison,

- the Pearson chi-square : $S_N = \sum_{j=0}^{N} \frac{(O_j - n(\alpha_{j+1} - \alpha_j))^2}{n(\alpha_{j+1} - \alpha_j)} \underset{H0}{\overset{d}{\sim}} \chi^2_N$

  and two of its possible modifications :
- the Nass test
- the LR (asymptotic Likelihood Ratio test)

▷ Multi-steps experiment : Static view

- Simulate multinomial data where F is normal (benchmark) and G of various types : $t5$, $t3$ and skewed $t3$
- Count the simulated observations lying between the N quantiles of G, where $N = 1, 2, 4, 8, 16, 32, 64$
- Choose different lengths n_1 for the sample of backtest, namely $n_1 = 250, 500, 1000, 2000$, and estimate the rejection probability for the null hypothesis (H0) using 10 000 replications (changing seeds)
- Why introducing several quantiles (ES) ?

	$VaR_{0.975}$	$VaR_{0.99}$	Δ_1	$ES_{0.975}$	Δ_2
Normal	1.96	2.33	0.00	2.34	0.00
t5	1.99	2.61	12.04	2.73	16.68
t3	1.84	2.62	12.69	2.91	24.46
st3 ($\gamma = 1.2$)	2.04	2.99	28.68	3.35	43.11

TABLE: Values of $VaR_{0.975}$, $VaR_{0.99}$ and $ES_{0.975}$ for four distributions (mean0, var 1) used in simulation study (Normal, Student t5, Student t3, skewed Student t3 with skewness parameter $\gamma = 1.2$). Δ_1 column shows percentage increase in $VaR_{0.99}$ compared with normal distribution ; Δ_2 column shows percentage increase in $ES_{0.975}$ compared with normal distribution.

TABLE: Rejection rate for the null hypothesis (H0) on a sample size of length n_1, using a multinomial approach with 3 possible tests (χ^2, Nass, LR) to backtest simultaneously the $N = 2^k$, $1 \leq k \leq 6$, quantiles VaR_{α_j}, $1 \leq j \leq N$, with $\alpha_1 = \alpha = 97.5\%$, on data simulated from various distributions (normal, Student $t3$, $t5$ and skewed $t3$)

	test	Pearson							Nass							LRT						
G	$n \mid N$	1	2	4	8	16	32	64	1	2	4	8	16	32	64	1	2	4	8	16	32	64
Normal	250	3.9	4.7	5.6	8.5	10.5	14.1	21.5	3.9	3.5	5.0	4.7	5.1	5.0	4.8	7.5	10.0	6.5	6.5	6.5	6.2	6.1
	500	3.9	4.4	5.2	6.6	8.6	12.3	16.2	3.9	3.9	4.7	4.7	5.5	5.5	5.3	5.9	5.8	5.5	5.6	5.3	5.3	5.2
	1000	5.0	5.2	5.0	5.6	7.2	9.0	12.0	5.0	4.8	4.7	4.9	5.1	5.3	5.1	4.1	5.5	5.5	5.8	5.6	5.6	5.7
	2000	5.0	4.5	4.8	5.0	6.3	7.2	8.8	5.0	4.3	4.5	4.5	5.3	5.1	4.9	4.2	4.9	4.7	5.0	5.1	5.1	5.0
t5	250	4.1	10.2	14.1	20.8	22.4	27.0	34.2	4.1	7.7	12.8	14.1	13.4	14.4	13.0	6.9	14.4	15.8	21.6	26.6	30.7	33.7
	500	5.2	15.7	22.1	28.4	32.2	36.2	39.8	5.2	14.3	20.5	24.5	26.6	26.0	22.7	6.5	15.5	26.9	36.6	44.7	50.4	54.8
	1000	6.9	26.7	40.2	48.2	53.0	54.8	55.8	6.9	25.5	39.5	46.2	48.6	47.7	43.8	5.2	26.1	46.4	61.8	71.4	76.7	80.5
	2000	7.3	47.2	70.4	79.3	82.5	82.8	82.0	7.3	47.0	69.6	78.2	80.8	80.2	77.0	5.8	48.0	77.4	89.5	94.4	96.6	97.6
t3	250	3.6	7.3	13.7	21.1	19.4	25.8	28.1	3.6	5.6	12.1	14.8	13.4	13.2	13.6	10.3	24.4	24.4	35.4	43.2	48.0	51.9
	500	4.8	16.1	25.2	32.7	35.2	40.1	38.6	4.8	15.5	22.4	28.7	32.3	29.4	26.4	9.5	26.2	44.2	58.6	67.9	73.8	78.0
	1000	9.9	37.4	55.6	62.9	65.2	64.8	64.2	9.9	35.2	54.1	60.3	61.4	59.9	54.7	9.7	47.2	75.4	87.7	93.2	95.5	96.8
	2000	16.6	73.1	91.0	94.5	94.9	93.9	92.1	16.6	72.7	90.5	94.2	94.3	92.6	89.6	16.5	79.5	96.8	99.4	99.8	99.9	100.0
st3	250	5.4	18.9	28.8	40.0	38.7	46.3	50.5	5.4	15.3	26.3	30.5	30.2	30.5	30.7	8.0	24.6	33.5	46.5	55.1	60.8	65.4
	500	6.9	34.9	50.7	60.6	64.6	69.5	70.2	6.9	33.2	47.6	56.2	61.4	60.0	56.8	7.9	35.9	59.3	73.6	81.6	86.2	88.9
	1000	9.5	62.3	83.0	89.1	91.3	92.1	92.0	9.5	61.4	82.3	88.1	90.0	90.0	87.9	6.9	62.3	88.1	95.3	97.9	98.9	99.2
	2000	12.2	90.7	98.7	99.7	99.8	99.8	99.7	12.2	90.7	98.6	99.7	99.7	99.7	99.5	9.8	91.6	99.3	99.9	100.0	100.0	100.0

Synopsis for the static view

- For all non normal distributions, considering only the VaR (1 point) does not reject the normal hypothesis, for all tests. The VaR does not capture enough the heaviness of the tail. Taking 2 quantiles improves slightly but not enough. The mulinomial approach with $N \geq 4$ gives certainly much better results than the traditional binomial backtest
- The heavier the tail of the tested distribution, the more powerful is the multinomial test
- For all the distributions, increasing the number n_1 of observations improves the power of all tests
- The Nass test with $N = 4$ or 8 seems to be a good compromise between an acceptable size and power and to be slightly preferable to the Pearson test with $N = 4$.
- In comparison with Nass, the LRT with $N = 4$ or $N = 8$ is a little oversized but very powerful.
- If obtaining power to reject bad models is the overriding concern, then the LRT with $N > 8$ is extremely effective

Determining an 'optimal' N, s.t. N the smallest possible to provide a combination of reasonable size and power of the backtest (to have a backtest comparable with the one of the VaR in terms of simplicity and speed of procedure) :

- Select N s.t. the size of the 3 corresponding tests lies below 6%.
- For $n_1 \geq 500$, the size varies between 4.2% and our threshold 6%. For the first two tests (chi-square and Nass), the size increases with N, whereas, for the LRT, it is more or less stable (slightly nonincreasing with increasing N)
- The power increases with N and the sample size n_1, for the 3 tests. It makes sense : the more information we have in the tail, the easier it is to distinguish between light and heavy tails

$\hookrightarrow$ $N = 4$ or 8 : overall reasonable choice.

G	n | test	Bin (0.99)	Pearson (4)	Nass (4)	LRT (4)	LRT (8)
Normal	250	4.0	5.6	5.0	6.5	6.5
	500	3.7	5.2	4.7	5.5	5.6
	1000	3.8	5.0	4.7	5.5	5.8
	2000	5.4	4.8	4.5	4.7	5.0
t5	250	17.7	14.1	12.8	15.8	21.6
	500	22.4	22.1	20.5	26.9	36.6
	1000	33.0	40.2	39.5	46.4	61.8
	2000	59.9	70.4	69.6	77.4	89.5
t3	250	13.5	13.7	12.1	24.4	35.4
	500	16.2	25.2	22.4	44.2	58.6
	1000	22.3	55.6	54.1	75.4	87.7
	2000	41.4	91.0	90.5	96.8	99.4
st3	250	31.2	28.8	26.3	33.5	46.5
	500	44.2	50.7	47.6	59.3	73.6
	1000	66.2	83.0	82.3	88.1	95.3
	2000	92.9	98.7	98.6	99.3	99.9

Table 4: Comparison of estimated size and power of one-sided binomial score test with $\alpha = 0.99$ and Pearson, Nass and likelihood-ratio test with $N = 4$ and LRT with $N = 8$. Results are based on 10000 replications

Results from binomial tests are much more sensitive to the choice of α. We have seen before that their performance for $\alpha = 0.975$ is very poor. The multinomial tests using a range of thresholds are much less sensitive to the exact choice of these thresholds, which makes them a more reliable type of test.

Other experimental design (static view)

The style of backtest is designed to mimic the procedure used in practice where models are continually updated to use the latest market data. We assume that the estimated model is updated every 10 steps.

In each experiment we generate a total dataset of $n + n_2$ values from the true distribution G; we use the same four choices as in the previous section. The length n of the backtest is fixed at the value 1000.

The modeller uses a rolling window of n_2 values to obtain an estimated distribution F, with n_2=250, 500 respectively. We consider 4 possibilities for F :

The oracle who knows the correct distribution and its exact parameter values.

The good modeller who estimates the correct type of distribution (normal when G is normal, Student t when G is t5 or t3, skewed Student when G is st3).

The poor modeller who always estimates a normal distribution (which is satisfactory only when G is normal).

The industry modeller who uses the empirical distribution function by forming standard empirical quantile estimates, a method known as *historical simulation* in industry.

We consider the same three multinomial tests as before and the same numbers of levels N. The experiment is repeated 1000 times to determine rejection rates.

Results :

- Again clear that taking values of $N \geq 4$ gives reliable results, superior to those obtained when $N = 1$ or $N = 2$.
- The use of only one or two quantile estimates does not seem sufficient
  - ↪ to discriminate between light and heavy tails
  - ↪ a fortiori to construct an implicit backtest of ES based on N VaR levels.

▷ Multi-steps experiment : Dynamic view

Backtesting experiment conducted now in a time-series setup. The true data-generating mechanism for the losses is a stationary GARCH(1,1) model with Student innovations (parameters chosen by fitting this model to S&P index log-returns for the period 2000–2012 (3389 values)).
A variety of forecasters use different methods to estimate the conditional distribution of the losses at each time point and deliver VaR estimates.
Length of the backtest : $n = 1000$ (approximately 4 years)
Each forecaster uses a rolling window of n_2 values to make their forecasts.
We consider the values $n_2 = 500, 1000$ respectively (longer window lengths than in the static backtest study since more data is generally needed to estimate a GARCH model reliably). All models are re-estimated every 10 time steps. *Experiment repeated* 500 *times to determine rejection rates for each forecaster.*

Description

Oracle : the forecaster knows the correct model and its exact parameter values.

GARCH.t : the forecaster estimates the correct type of model (GARCH(1,1)- t).

GARCH.HS : the forecaster uses a GARCH(1,1) model to estimate the dynamics of the losses, but applies empirical quantile estimation to the residuals to estimate quantiles of the innovation distribution and hence quantiles of the conditional loss distribution (method called 'filtered historical simulation'). We have already noted in the static backtesting experiment that empirical methods are only acceptable when we use a sufficient quantity of data.

GARCH.EVT : the forecaster uses a variant on GARCH.HS in which an EVT tail model is used to get slightly more accurate estimates of conditional quantiles in small samples.

GARCH.norm : the forecaster estimates a GARCH(1,1) model with normal innovation distribution.

ARCH.t : the forecaster misspecifies the dynamics of the losses by choosing an ARCH(1) model but correctly guesses that the innovations are t-distributed.

ARCH.norm : as in GARCH.norm but the forecaster misspecifies the dynamics to be ARCH(1).

HS : the forecaster applies standard empirical quantile estimation to the data. As well as completely neglecting the dynamics of market losses, this method is prone to the drawbacks of empirical quantile estimation in small samples.

TABLE: Estimated size and power of three different types of multinomial test (Pearson, Nass, likelihood-ratio test (LRT)) based on exceptions of N levels. Results are based on 500 replications of backtests of length 1000

	Test	Pearson							Nass							LRT						
n_2	$F \mid N$	1	2	4	8	16	32	64	1	2	4	8	16	32	64	1	2	4	8	16	32	64
500	Oracle	6.0	4.0	3.8	5.0	5.6	9.6	9.4	6.0	3.2	3.6	4.2	2.8	4.8	4.0	3.4	4.8	5.2	5.0	5.4	5.6	5.6
	GARCH.t	6.8	5.6	6.2	8.0	8.2	12.8	18.4	6.8	5.0	6.0	6.2	7.0	7.6	6.4	4.6	5.0	5.4	4.8	4.6	5.2	5.2
	GARCH.HS	1.6	1.6	4.4	11.8	25.4	92.0	98.8	1.6	1.4	4.4	10.8	20.4	85.0	97.4	0.8	1.6	3.6	2.0	2.0	5.6	13.0
	GARCH.EVT	2.2	3.6	3.6	7.2	7.6	12.2	16.2	2.2	3.6	3.2	6.0	5.0	6.8	7.4	0.8	3.6	2.0	0.8	1.2	1.0	1.0
	GARCH.norm	10.8	34.0	50.4	61.6	66.0	68.6	69.4	10.8	32.2	49.4	60.0	61.4	63.2	55.4	8.2	34.0	55.2	71.2	79.8	85.0	87.4
	ARCH.t	34.0	32.4	32.0	29.8	29.4	33.8	39.6	34.0	31.4	31.2	28.6	26.8	25.6	28.0	30.4	31.2	31.4	31.6	31.8	31.8	31.2
	ARCH.norm	96.2	99.6	99.6	99.8	100.0	100.0	100.0	96.2	99.6	99.6	99.8	100.0	100.0	100.0	95.0	99.6	99.6	99.8	100.0	100.0	100.0
	HS	39.4	38.8	39.8	42.2	49.8	80.2	90.0	39.4	38.6	39.8	40.8	48.0	77.0	85.0	36.8	40.0	44.8	43.8	42.2	49.2	55.8
1000	Oracle	4.2	3.4	3.8	3.4	5.0	7.6	10.2	4.2	3.2	3.8	2.8	3.6	3.8	3.8	3.4	2.6	3.2	2.6	2.6	2.4	2.4
	GARCH.t	5.8	4.6	6.2	5.2	6.0	11.2	12.8	5.8	3.8	5.2	3.2	4.2	6.6	6.6	4.4	2.8	3.6	3.6	3.4	4.8	4.0
	GARCH.HS	3.0	2.0	2.6	4.4	10.2	21.2	69.0	3.0	1.8	2.2	4.0	7.2	13.2	56.0	1.8	1.6	2.6	3.4	3.8	3.0	5.2
	GARCH.EVT	2.6	3.4	4.2	4.2	7.0	7.2	10.4	2.6	3.4	3.6	3.4	5.0	3.8	4.2	1.6	4.6	3.2	2.6	2.0	1.8	1.8
	GARCH.norm	9.4	30.6	45.6	52.2	58.4	61.4	63.6	9.4	29.8	44.6	49.6	53.0	54.6	50.6	6.4	28.4	49.8	65.2	76.6	83.0	86.6
	ARCH.t	42.4	36.8	32.8	28.0	25.0	30.2	33.8	42.4	36.0	32.2	27.0	23.0	25.6	27.2	39.4	40.2	39.6	40.0	40.2	40.4	40.8
	ARCH.norm	82.8	94.6	97.6	98.2	98.6	98.2	98.6	82.8	94.4	97.6	98.0	98.2	98.0	97.8	80.8	95.2	98.8	98.8	99.2	99.2	99.4
	HS	51.4	51.0	45.0	37.2	34.6	39.8	55.6	51.4	50.6	44.4	35.0	31.8	36.2	49.8	49.2	51.8	52.6	53.8	53.8	53.0	55.4

In conclusion, this experiment confirms that using $N = 4$ or 8 quantiles gives an effective multinomial test ; $N = 4$ is appropriate if using a Pearson or Nass tests and $N = 8$ gives superior power if using the LRT.

A procedure to implicitely backtest ES

▷ In view of the numerical results we can suggest an 'optimal' multinomial test.

- *Number of quantiles taken at intervals such that* $\mathbb{E}(O_j)$ *is constant ; it turns out that choosing* $N = 4$ *seems adequate, and* $N = 8$ *in LRT is the most powerful.*
- *Among the 3 possible tests, the Nass test and the LRT share on average the best results, taking into account both the test size and power, the Nass for the static view and the LRT for the dynamic one. The LRT is in general the most powerful and might be used if we want more sensitivity, in particular w.r.t. the parameters.*

▷ The ES estimated with a model that is not rejected by our multinomial test, is implicitely accepted by our backtest. Hence we can use the same rejection criterion for ES as for the null hypothesis (H0) in the multinomial test.

▷ A traffic light system has been proposed recently by the Basel Committee for Banking Supervision (Jan.2016) based on the backtest of two quantiles, to improve the binomial approach for one quantile. We can use a similar metaphor to illustrate the decision criterion about validating or not the ES estimate, and compare it to the binomial ($N = 1$) or $N = 2$ approaches.

Conclusion

- We developed several variants of a multinomial test to simultaneously judge the backtesting performance of trading book models at different VaR levels ; it gives then an implicit backtest for ES.
- Evaluation of this multinomial approach in a series of Monte Carlo simulation studies of size and power, and further experiments that replicate typical conditions of an industry backtest. It aims as understanding better the test itself and set a benchmark ; it has been carried out on real data (one example on S&P500 ; see the paper)
- The multinomial test distinguishes much better between good and bad models (particularly in longer backtests) than :
  - the standard binomial exception test
  - a multinomial test based on two quantiles

- Backtesting simultaneously 4 or 8 (for LRT) quantile levels seems an optimal choice whatever is the test in terms of balancing simplicity and reasonable size and power properties
- This multinomial backtest could be used for ES as a regular routine, as done usually for the VaR with the binomial backtest, giving even more arguments to move from VaR to ES in the future Basel III.
- Possible to design a traffic-light system for the application of capital multipliers and the imposition of regulatory interventions, completely analogous to the current traffic-light system based on VaR exceptions over a 250 day period at the 99% level.
- We would suggest moving to longer backtesting periods than 250 days to obtain more powerful discrimination between good and bad backtesting results.
- For sharper results, other backtests may complement this one, as the PIT already used for distribution forecasts, or methods based on realized p-values, or e.g. joint testing procedures of expected shortfall and VaR proposed by Acerbi and Székely

Main references for this study :

BCBS (2016). *Standards. Minimum capital requirements for market risk*. Basel Committee on Banking Supervision, January 2016.

Y. CAI, K. KRISHNAMOORTHY (2006). Exact size and power properties of five tests for multinomial proportions. *Comm. Statistics - Simulation and Computation* **35(1)**, 149-160.

S.D. CAMPBELL (2006). A review of Backtesting and Backtesting Procedures. *Journal of Risk* **9(2)**, 1-17.

S. EMMER, M. KRATZ, D. TASCHE (2015). What is the best risk measure in practice ? A comparison of standard measures. *Journal of Risk* **18**, 31-60.

M. KRATZ, Y. LOK, A. MCNEIL (2016). A multinomial test to discriminate between models. (*ASTIN 2016 proceedings* and *preprint*)

A. MCNEIL, R.FREY, P. EMBRECHTS (2015, 2nd Ed.). Quantitative Risk Management. *Princeton Univ. Press*

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 318984 - **RARE** (**R**isk **A**nalysis, **R**uin theory, **E**xtremes)