An implicit backtest for ES via a simple multinomial approach

Marie KRATZ

ESSEC Business School
Paris Singapore

CREAR
ESSEC Research Center on Risk

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

Vth IBERIAN CONGRESS OF ACTUARIES, Lisboa 2016
ISEG, June6-7, 2016

- 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 :
    1. Solvency Capital Requirements (SCR) of Solvency II : VaR (99.5% yearly)
    2. Target capital for the Swiss Solvency Test : ES (99% yearly)
    3. Basel II : VaR (99% daily)
    4. 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 if the violation process $I_t(\alpha)$ satisfies the two conditions (Christoffersen, 03) :

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

- 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 fact that $F(X) \overset{d}{=} \mathcal{U}(0,1)$)

(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 ?
- 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, assuming :
  - mean and variance of the distributions match, to focus on misspecification of kurtosis and skewness
  - we might be subject to estimation error, as in practice.
- 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 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, with $G \neq F$, where the parameters are given by

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

Ex : 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. $\beta_j = F(G^{\leftarrow}(\alpha_j)) \neq \alpha_j$.

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

Here we use : 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 and the LR (asymptotic Likelihood Ratio) tests, for comparison.

Careful when using the LRT, as it cannot be used with an unrestricted alternative hypothesis (because it could lead to to an undefined test statistic when there are no observation in some of the cells). We need in such a case a parametric form. In our applications, we consider the alternative hypothesis (H1) such that the cell probabilities are based on a normal distribution with two parameters μ_1 and σ_1, where $\mu_1 \neq \mu_0 = 0$ and $\sigma_1 \neq \sigma_0 = 1$.

▷ 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)
- Two cases :

  (i) mean and variance of the benchmark normal data match the ones of the fitted model

  (ii) mean and variance are estimated, involving estimation errors

- Additional question for (ii) : how much data for the estimation of the models to make sure size an power remain reasonable ? We will try $n_2 = 250, 1000, 2000$.

Example when no parameter estimation error

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 $t3$)

		Chi Square							Nass							LR						
	n1	1	2	4	8	16	32	64	1	2	4	8	16	32	64	1	2	4	8	16	32	64
Standard Normal	250	4,1%	5,0%	5,3%	9,1%	11,3%	15,0%	22,3%	4,1%	3,7%	5,0%	5,3%	5,6%	5,3%	5,2%	7,7%	10,4%	6,3%	6,2%	6,4%	6,2%	6,0%
	500	4,4%	4,8%	5,2%	6,2%	8,4%	11,8%	15,7%	4,4%	4,2%	4,6%	4,6%	5,4%	5,4%	5,0%	6,5%	5,9%	5,7%	5,7%	5,5%	5,5%	5,4%
	1000	5,1%	4,6%	5,2%	5,9%	7,4%	9,2%	12,3%	5,1%	4,2%	4,9%	5,1%	5,3%	5,4%	5,4%	4,2%	5,3%	5,5%	5,4%	5,2%	5,2%	5,2%
	2000	5,2%	4,9%	5,0%	5,7%	6,3%	7,4%	9,7%	5,2%	4,8%	4,6%	5,1%	5,4%	5,3%	5,6%	4,3%	5,3%	5,0%	4,9%	4,9%	4,9%	4,9%
t5	250	5,0%	10,6%	14,5%	21,9%	23,1%	27,4%	34,0%	5,0%	8,3%	13,2%	14,8%	14,3%	14,8%	13,7%	8,0%	15,6%	16,6%	22,6%	27,0%	31,1%	34,1%
	500	5,4%	15,8%	22,1%	28,5%	31,8%	36,1%	39,0%	5,4%	14,5%	20,1%	24,4%	26,4%	25,4%	22,7%	6,8%	16,0%	26,6%	36,9%	44,7%	50,3%	54,5%
	1000	6,6%	27,5%	41,7%	49,8%	54,1%	55,0%	55,6%	6,6%	26,4%	40,9%	47,2%	49,9%	48,0%	43,7%	5,0%	26,9%	48,3%	63,0%	72,4%	78,3%	81,4%
	2000	7,5%	47,9%	71,0%	79,7%	82,7%	82,8%	81,6%	7,5%	47,8%	70,2%	78,7%	81,2%	79,8%	76,7%	6,0%	48,9%	77,4%	89,7%	94,5%	96,7%	97,8%
t3	250	3,8%	7,1%	13,3%	20,8%	19,5%	25,6%	28,3%	3,8%	5,3%	11,7%	14,2%	13,8%	13,8%	13,9%	10,3%	24,7%	24,3%	35,6%	42,2%	48,1%	52,1%
	500	5,3%	16,0%	24,3%	32,5%	34,4%	39,6%	38,5%	5,3%	15,4%	21,4%	27,8%	31,6%	28,9%	25,8%	9,8%	27,1%	44,7%	58,8%	68,1%	73,9%	77,7%
	1000	9,9%	37,7%	56,5%	63,6%	65,4%	64,4%	64,4%	9,9%	35,6%	55,2%	60,9%	62,0%	60,0%	54,3%	9,8%	47,6%	75,3%	88,0%	93,0%	95,6%	96,6%
	2000	17,4%	73,7%	90,9%	94,6%	94,8%	93,7%	91,9%	17,4%	73,3%	90,4%	94,0%	94,2%	92,5%	89,6%	17,4%	80,3%	96,7%	99,3%	99,8%	100,0%	100,0%
skt3	250	13,6%	38,7%	52,0%	64,0%	63,5%	69,8%	73,8%	13,6%	34,2%	49,5%	54,6%	55,0%	55,2%	55,0%	14,3%	34,8%	53,5%	66,3%	73,9%	78,4%	81,6%
	500	24,0%	63,3%	79,0%	85,7%	88,1%	89,8%	90,7%	24,0%	60,8%	77,2%	82,9%	86,0%	85,2%	84,1%	24,1%	58,6%	81,7%	90,7%	94,4%	96,4%	97,2%
	1000	41,7%	89,4%	97,1%	98,7%	99,2%	99,3%	99,3%	41,7%	89,0%	96,9%	98,6%	99,1%	99,0%	98,8%	35,2%	87,5%	97,9%	99,6%	99,8%	100,0%	100,0%
	2000	66,2%	99,6%	100,0%	100,0%	100,0%	100,0%	100,0%	66,2%	99,6%	100,0%	100,0%	100,0%	100,0%	100,0%	61,6%	99,5%	100,0%	100,0%	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. The mulinomial approach 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 LR test seems to be the most powerful and the Nass the less one
- The LR test is very sensitive to the estimation error, due to $(H1)$
- In general, taking $n_1 = 250$ does not provide satisfactory results, so we will not base our discussion on this sample size.

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.

▷ Dynamic view

Numerical application : we devise a multisteps experiment to see how the multinomial test performs :

1. Generate a sample data path of length 3000 using a GARCH(1,1) model with student-t innovations (our benchmark model)

2. Tested models (using a rolling window size of 1000) :
   - GARCH(1,1) model with student-t innovations
   - GARCH(1,1) model with standard normal innovations
   - GARCH(1,1) model with historical simulation method applied to the residuals (i.e. dynamic historical simulation method)
   - ARCH(1) model with student-t innovations
   - ARCH(1) model with standard normal innovations
   - Historical simulation method.

3. Backtest the obtained sets of VaR_{u_j} using the multinomial test.

4. Repeat step 1 to 3 500 times to estimate the rejection rate of each test.

TABLE: Rejection rate of the χ^2 goodness of fit test, with $\kappa = 97.5\%$, $N = 1, 4, 8, 16$

Model	N=1	4	8	16
GARCH-t (benchmark)	2.8%	4.0%	4.0%	6.6%
GARCH HS	0.8%	1.2%	2.0%	1.8%
ARCH-t	36.4%	35.0%	31.6%	30.8%
HS	42.2%	47.6%	43.4%	43.0%
GARCH normal	13.4%	69.6%	76.0%	79.8%
ARCH normal	75.4%	100.0%	100.0%	100.0%

- Reasonable size whenever $N \leq 8$
- The GARCH HS is not rejected as we would expect (since very close to the benchmark model). The HS method aplied to the innovations gives naturally a good approximation of the Student innovations
- the multinomial test accepts when the tails are treated correctly and strongly rejects the wrong models
- this test discriminates better the tails of the models than the types respectively, having the same tail or being HS model
- it is a very powerful test for both wrong model and innovation assumptions
- Compared to the Binomial test, the χ^2 test has a much higher power in detecting misspecification in the innovation assumption of the predictive distribution
- Size and the power of the χ^2-test leads to select $N = 4$ or 8. For $N = 4$, the model assumption is more discriminated than the innovation assumption ; for $N = 8$, reverse (makes sense as we would consider more points in the tail).

Conclusion

- We developed a multinomial test to discriminate between models ; it gives an implicit backtest for ES.
- Evaluation of this approach on simulated data ; it has been carried out on real data (preprint on arXiv soon)
- The multinomial test distinguishes much better between good and bad models, than :
  - the standard binomial exception test
  - a multinomial test based on two quantiles, as suggested in Basel 2016
- Backtesting simultaneously 4 quantiles seems an optimal choice in terms of simplicity and speed of the procedure, as well as in terms of reasonable size and power of the backtest.
- 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.
- For sharper results, other backtests may complement this one, as the PIT already used for distribution forecasts, or more recent ones (e.g. 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.

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)