A multinomial test to discriminate between models

Marie Kratz[*], Yen H. Lok [†], Alexander McNeil [‡]

Abstract

Replacing Value-at-Risk (VaR) by Expected Shortfall (ES) in Basel 3 is under current discussion, as ES is in general a better risk measure than VaR, more reliable tool for risk management. Hence the question of providing a backtest for ES, as handy in practice as the popular binomial backtest based on a violation process, used for the VaR. It is what we propose in this study. Following the idea by Emmer *et al.* of considering an empirical approach that consists in replacing ES by a set of a small number of quantiles for the backtesting, comes the natural proposition of a simple multinomial backtest for ES. It turns out to give reasonable results, certainly much better than with the binomial backtest, helping to distinguish between models.

Keywords: backtesting; coherence; expected shortfall; model validation; risk measure; risk management;

1 Introduction

Much literature has been devoted to propose and study risk measures, as they constitute an essential tool in terms of risk management and model validation. Without reducing the rich literature on this topic, it is worth mentioning a few studies with a big impact in practice. We can start with the work by Artzner *et al.* ([2]) who formalize mathematically the properties that should be expected from a risk measure, defining the coherence, one of its axioms being the subadditivity. If the subadditivity is clearly needed to be able to measure the diversification benefit of a risk portfolio, another big issue at the heart of a risk strategy, is about deciding the capital allocation among the various risks of the portfolio. Here, Tasche ([14], [15], [16]) provides an optimal solution in terms of capital allocation for coherent risk measures. For a given choice of risk measure, the next question is then to evaluate it properly, and to check if the realized losses, observed ex post, are in line with the ex ante forecasts. The set of statistical procedures designed to compare realizations with forecasts is known under the name of backtesting. It is an inescapable step in model validation, as emphasized by the new regulation.

The two main risk measures used so far in financial institutions and regulation are the Value-at-Risk (VaR) and Expected Shortfall (ES; named also Tail Value-at-Risk, TVaR), with a domination of the VaR in banks and regulation. The main usage of these risk measures is to compute, from the probability distribution of the firm's value, the Risk Adjusted Capital in its different forms: Solvency Capital Requirements (SCR) of Solvency II: yearly VaR(99.5%), SCR for the Swiss Solvency Test: yearly ES(99%), Basel II: daily VaR(99%).

[*] ESSEC Business School, CREAR risk research center; E-mail: kratz@essec.edu

[†] Heriot Watt University; E-mail: yhl30@hw.ac.uk

[‡] Heriot Watt University; E-mail: A.J.McNeil@hw.ac.uk

Although Expected Shortfall (ES) is in general a better risk measure than Value-at-Risk (VaR) because of its mathematical properties (in particular its coherence), it has also been proved by Gneiting ([10]) to be non elicitable, leading to less straightforward backtesting methods than, e.g., for VaR. It gave rise to a debate, but, as pointed out recently by Acerbi & Székely ([1]), elicitability (or lack of elicitability) is not relevant for backtesting of risk measures but rather for comparing the forecast performance of different estimation methods.

In practice, in order to assess the quality of the VaR predictions, a popular and direct procedure has been proposed, namely a binomial test on the proportion of violations (see [7], [6]). Under the hypothesis of independence between the components of the violation process $(I_t(\alpha))$ defined by $I_t(\alpha) = \mathbf{1}_{\left\{L(t)>VaR_\alpha(L(t))\right\}}$, replacing VaR by its estimates, we test if this estimated process behaves like independent and identically distributed Bernoulli random variables with violation (success) probability close to $1-\alpha$.

This paper is a prelude to a more developed study on the construction of a multiomial backtest for ES, as simple and costless as it is for the VaR, in order to get a regular backtest routine, as it is done usually for the VaR with the binomial backtest.

For this construction, we proceed via an approximation of ES and introduce a multinomial approach. The idea came naturally from the following approximation of ES proposed by Emmer *et al.* ([9]:

$$\mathrm{ES}_\alpha(L) \;\approx\; \frac{1}{4}\,[\,q(\alpha)+q(0.75\,\alpha+0.25)+q(0.5\,\alpha+0.5)+q(0.25\,\alpha+0.75)\,] \tag{1.1}$$

where $q(\gamma) = VaR_\gamma(L)$. Hence, if the four $q(a\alpha+b)$ are successfully backtested, then also the estimate of $\mathrm{ES}_\alpha(L)$ might be considered reliable. We can then build a procedure based on simultaneously backtesting multiple VaR estimates evaluated within the same method as the one used to compute the ES estimate.

Note that 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 (see [3]). However we will see that using only one quantile after the VaR is still not enough to provide a good test.

In this paper we examine the relevance of this type of approximations (1.1), and so, of the multinomial test, answering the following 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? looking at three possible forms of the multinomial test, is it one that would be more recommandable? to help us answering these questions, we proceed to two experiments, and compute for each, size and power of the corresponding multinomial test. First we take a static view and test distributional forms that might be typical for the trading book. We check if the multinomial test distinguishes well between them, in particular between their tails. Then, we turn to a 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.

2 Multinomial Test

Suppose we have a series of observed loss $(L_t, t = 1, \ldots, n)$, and a predictive model that produces the corresponding ES_α forecast (at threshold α). We wish to test whether the predictive model is adequate in modelling the tail distribution of the loss.

In [9], it has been suggested to approximate ES via (1.1), but we can investigate how many and which quantiles would be needed for a better model validation, varying the number of quantiles in (1.1).

Let N be the number of quantiles that will be used to approximate ES, namely

$$\mathrm{ES}_\alpha(L) \simeq \frac{1}{N}\sum_{j=1}^{N} q_j \quad \text{where } q_j = q(\alpha_j) = VaR_{\alpha_j}(L),\ \alpha_1 := \alpha < \ldots < \alpha_N < \alpha_{N+1} := 100\%\,. \tag{2.1}$$

Then we build a multinomial test which tests **simultaneously** the N VaR's, introducing, as for the VaR backtest (see [7], [6]), the so-called violation process $(I_t(\alpha))_t$ of VaR as

$$I_t(\alpha) = \mathbf{1}_{\left(L_t > VaR_\alpha(L_t)\right)} \tag{2.2}$$

$\mathbf{1}$ denoting the indicator function and L_t the loss at time t.

Forecasting the VaR at time $t+1$ by the VaR at time t:

$$\widehat{q}_{\alpha,t+1} := \widehat{VaR}_\alpha(L(t+1)) := VaR_\alpha(L(t)),$$

Christoffersen ([6]) showed that VaR forecasts are valid if and only if the violation process $I_t(\alpha)$ satisfies two conditions:

- the unconditional coverage hypothesis : $\mathrm{E}\left[I_t(\alpha)\right] = 1-\alpha$, and
- the independence condition: $I_t(\alpha)$ and $I_s(\alpha)$ are independent for $s \neq t$

under which the number of violations has a binomial distribution with success (violation) probability $1-\alpha$.

Here, testing simultaneously N VaR's (with $N > 1$) then leads to a multinomial distribution (under the same assumptions) and we can set the null hypothesis of the multinomial test as

$$(\mathrm{H0}): \quad p_j := \mathrm{E}\left[1_{(L_t > \widehat{q}_{j,t})}\right] (= \mathrm{P}\left[L_t > \widehat{q}_{j,t}\right]) \;=\; p_{j,0} := 1-\alpha_j, \quad \forall\, j = 1,\cdots,N\,. \tag{2.3}$$

to validate the VaR forecasts, then the ES by approximation. To judge the relevance of this test, we compute its size $\gamma = \mathrm{P}\,(\text{reject H0})|\text{H0 true}]$ (type I error) and its power $1-\beta = 1-\mathrm{P}\left[(\text{accept H0})|\text{H0 wrong}\right]$ (1- type II error).

Various test statistics can be used to describe the event (reject of H0). In [4], Cai and Krishnamoorthy provided a relevant numerical study of the properties of five possible tests for testing the multinomial proportions. Here we propose to use three of them, quite standard, namely the Pearson chi-square ([13]), and two of its possible modifications, the Nass ([12]) and the LR (asymptotic Likelihood Ratio; see e.g. [5]) tests, for comparison.

The schema of the general procedure is as follows:

- We compute a series of $q_j = \mathrm{VaR}_{\alpha_j}$, $j = 1,\ldots,N$, with $\alpha_1 = \alpha < \ldots < \alpha_{N+1} = 1$. The choice of N depends on the available amount of data and will be chosen as $N = 2^k$, with $k = 0,1,\cdots,6$ to be discussed in terms of the obtained results.
- We count the number of observations in each quantile-based cell $(q_j; q_{j+1})$, introducing, for $j = 1,\ldots,N$,

  $$O_j = \sum_t 1_{(q_j < L_t < q_{j+1})}\,.$$

  For optimal performance of the multinomial test, we choose α_j such that $\mathrm{E}\,(O_j)$ is a constant, independent of j

- Replace VaR_{α_j} by its estimates
- Test simultaneously the VaR's, with null hypothesis (H0) defined in (2.3). The alternative hypothesis (H1) is an unrestricted one for the χ^2 and Nass tests. For the LRT, (H1) is based on the normal distribution with parameters μ_1 and σ_1^2 distinct from the mean μ_0 and variance σ_0^2 under the null
- Introduce the test statistic of your choice (the standard χ^2 statistic, or an improved approximation to it given by Nass, or another type defined as the LR statistics) for testing multinomial proportions
- Compute the size γ and the power $1-\beta$ of the test
- Comparison and discussion

Then we devise a multi-steps experiment on simulated data, with both static and dynamic views, to judge the performance of multinomial tests via their size and power, the ultimate goal being to design the best possible multinomial test for model validation.

3 Application on simulated data

To answer the main questions given in the introduction on the relevance of the multinomial approach we suggested, we consider two experiments. First, taking a static view, we test distributional forms that might be typical for the trading book, and see if the multinomial test distinguishes well between them, in particular between their tails. Then, we turn to a 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. We compare if this multinomial approach offers better results in terms of model validation than the binomial one, or that based on two quantiles only.

3.1 Static view

Here we consider typical distributions for the trading book, and look at the ability to distinguish between the distributional forms, in particular between their tails. We compute the size of the test, corresponding to the type I error assuming the forecaster works with the right returns distribution. Then, in the case he would work with the wrong model, we want to see if the test would detect it, so compute the test power.

We generate data from a standard normal distribution, as a benchmark, then from a variety of heavy-tailed and skewed distributions. We consider for instance Student distributions with 5 and 3 degrees of freedom ($t5$ and $t3$) to have moderate heavy and heavy tails respectively, and the skewed Student with 3 degrees of freedom (denoted sk$t3$).

Note that we assume that the mean and variance of the benchmark normal data match the ones of the fitted model, assuming they are known, in order to focus on misspecification of kurtosis and skewness.

We choose as null hypothesis (H0) that our data are drawn from a standard normal distribution, and define N quantile-based cells under this hypothesis, where $N = 2^k$, with $k = 0, 1, \cdots, 6$. The goal is then to use the multinomial approach to backtest simultaneously the N quantiles from various data which are normal, $t5$, $t3$ and sk$t3$ respectively, to calculate size and power of the test.

It means that for each data set, we count data in each cell and use the multinomial test to compare with expected numbers under the null hypothesis of normality.

We choose different lengths n_1 for the sample of backtest, namely $n_1 = 250, 500, 1000, 2000$.

We calculate then the size and the power for a backtest of length n_1, estimating the rejection probability for the null hypothesis (H0) using 10'000 replications, changing seeds. We present the obtained results in Table 1. The green color indicates good results ($< 6\%$ for the size; $\geq 70\%$ for the power); the red one, bad ones ($> 9\%$ for the size; $< 30\%$ for the power), and the dark red, very bad ones ($> 25\%$ for the size; $< 10\%$ for the power).

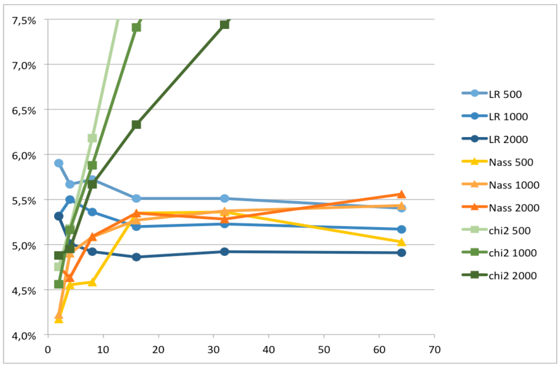
Table 1: Rejection rate for the null hypothesis (H0) on a sample size of length n1, using a multinomial approach with 3 possible tests (χ^2, Nass, LR) to backtest simultaneously the $N = 2^k$, $1 \leq k \leq 6$, quantiles VaR_{α_j}, $1 \leq j \leq N$, with $\alpha_1 = \alpha = 0.95$, on data simulated 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%

(i) Size of the tests

The size of the 3 tests introduced to judge the performance of the simultaneous backtest of the N VaR_{α_j} $(1 \leq j \leq N)$ can be read in the first row "Standard Normal" of Table 1. In Figure 1, we plot for each test, the size as a function of N; we do it for all the sample sizes we consider.

Figure 1: Size of the multinomial tests (χ^2 (in green), Nass (in yellow), LRT (in blue)) as a function of N, for sample sizes $n_1 = 500, 1000$ and 2000



Comments on Figure 1 :

- Results are more or less stable for Nass test and LRT, increasing slightly with N (for $N \geq 2$) for the Nass test and decreasing very slightly for the LRT from $N = 4$ (the difference is negligeable)

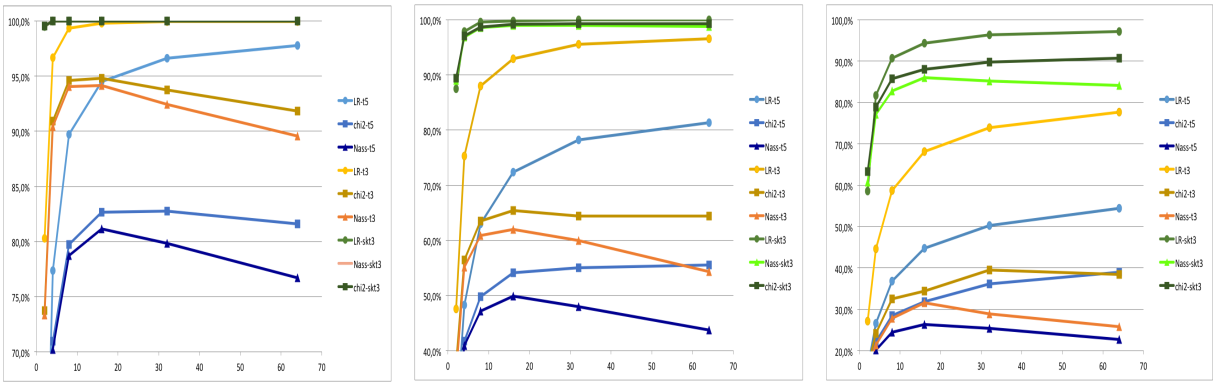
- For the χ^2-test, there are numerical problems to estimate the far extreme quantiles (with probability more or less 0) whenever $N \geq 16$

(ii) Power of the tests

In Figure 2, we present a set of three graphs of the power test results as a function of N for various models:

- On the left graph, we consider a Student $t5$. We know that by nature it is

Figure 2: Power of the multinomial tests (χ^2 (in green), Nass (in yellow), LRT (in blue)) as a function of N and for sample sizes $n_1 = 500, 1000$ and 2000. From left to right, samples come from t5, t3 and skewed t3 distributions, respectively.



more difficult to distinguish from the normal case, than when testing other distributions with heavier tails. We observe that:

- for the 3 tests, we need n_1 large enough ($n_1 = 2000$) to distinguish between normal and $t5$ distributions, in order to reject the normal hypothesis;
- for the LRT, it starts to reject the normal hypothesis at $n_1 = 1000$ but for large N (≥ 16), which makes sense as we need more information in the tail to be able to distinguish it from a normal one (as already noticed in the general comments).

- When considering on the middle graph a heavier tail, namely a Student $t3$, we observe that:
  - when comparing the power with that obtained for $t5$, it is larger, as expected. Results are more relevant;
  - to have a power larger than 70%, we need to take $n_1 = 2000$ for the χ^2 and Nass tests; for the LRT, we can consider $n_1 = 1000$ for $N \geq 4$, and $n_1 = 500$ for $N \geq 32$.
- On the right graph, we consider a skewed Student $t3$. Here, we obtain very powerful tests, even from $n_1 = 500$, whenever $N \geq 4$. This is due to the fact that the skewness pushes the tail on the right hand side.

To conclude, we can observe that:

- 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
- In general, taking $n_1 = 250$ does not provide satisfactory results, so we will not base our discussion on this sample size.

Let us look at the results to determine an 'optimal' N, such that N is the smallest possible to provide a combination of reasonable size and power of the backtest, in order to have a backtest comparable with the one of the VaR in terms of simplicity and speed of procedure.

- We consider the values of N such that 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 increaes 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 as the more information we have in the tail, the easier it will be to distinguish between light and heavy tails

Taking into account these observations on both the size and the power of the three tests, $N = 4$ or 8 seems an overall reasonable choice.

3.2 Dynamic view

Here we use a similar strategy as before to compute size and power of the multinomial test, choosing for instance the χ^2 one, and design the same experiment as for the static view, considering now a time series setup. We consider a Garch-t model with given parameters, as a benchmark model, and fit models of the various types: nonparametric (HS), or semiparametric (GARCH-HS) to avoid having to specify model fully, or parametric, in order to compare their test power.

Recall that the the standard unconditional historical-simulation method, named HS-method, can be thought of as estimating the distribution of the loss operator under the empirical distribution of the historical losses. It is the most popular method used by banks for the trading book.

So we generate a sample data path of length 3000 using a GARCH(1,1) model with Student-t innovations; it will be our benchmark model, denoted by GARCH-t in Table 2.

Then we fit to the benchmark sample, the following models, which aim at representing the various possible types of models (even if the list could, of course, be completed):

- a GARCH(1,1) model with residuals obtained by the standard unconditional historical-simulation (named HS) method; it will denoted by GARCH HS
- an ARCH(1) model with student-t innovations, denoted ARCH-t
- a model obtained by the HS method, denoted HS
- a GARCH(1,1) model with standard normal innovations, denoted GARCH normal
- an ARCH(1) model with standard normal innovations, denoted by ARCH normal

from which we deduce the respective $\text{VaR}^t_{\alpha_j}, j = 1, \ldots, N$, using a rolling window size of 1000.

The previous experiment (static view) has shown that N needs to be not too large to obtain a reasonable test size ($N = 4$ or 8), hence for this experiment, it seems reasonable to investigate the various cases with $N \leq 16$. Next, we backtest the obtained sets of VaR_{α_j} using our multinomial approach with the χ^2 test.

We estimate the rejection probability for the null hypothesis (H0) using 3500 replications.

In Table 2, we provide the obtained results when considering $\alpha = 95\%$, N taking the values 1, 4, 8 or 16. Recall that the column $N = 1$ corresponds to the binomial test performed to backtest VaR_α.

Table 2: Rejection rate of the χ^2 test, with $\alpha = 0.95$, N = 1, 4; 8, 16, performed on data simulated from historical (HS), or GARCH / ARCH with innovations chosen from 3 different distributions: empirical (HS), normal and Student (t).

	χ^2			
Model	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	63.4%	69.6%	76.0%	79.8%
ARCH normal	75.4%	100.0%	100.0%	100.0%

On Table 2, we observe that:

- we have a reasonable size whenever $N \leq 8$ (see the results indicated in green when fitting the right model, 2nd row of the table)
- The GARCH HS is not rejected as we would expect, since it is 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 when both the model and the innovation assumptions are wrong
- 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
- Taking into account both the size and the power of the χ^2-test leads to select N = 4 or 8. The value $N = 4$ would discriminate more the model assumption than the innovation assumption, whereas for $N = 8$, it would be reverse (which makes sense as we would consider more points in the tail).

4 Conclusion

In this study, we developed a multinomial approach to discriminate between models, and applied it on simulated data. In an extended version of this work, we complet the tests, and carry out the method on real recent data.

As expected, the multinomial test distinguishes much better between good and bad models, than the standard binomial exception test. Backtesting simultaneously four 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 it is 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 ([8]), or new ones, like the one suggested recently by Acerbi and Székely ([1]) relying on Monte-Carlo simulation.

References

[1] C. Acerbi, B. Székely (2014). Back-testing expected shortfall. *Risk*, Dec., 1-6.

[2] P. Artzner, F. Delbaen, J.-M. Eber, D. Heath (1999). Coherent measures of risks. *Mathematical Finance* **9**, 203-228.

[3] BCBS (2013). *Fundamental review of the trading book: A revised market risk framework.* Basel Committee on Banking Supervision, October 2013.

[4] 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.

[5] G. Casella, R. Berger *Statistical Inference* (2002). Duxbury Advanced Series (2nd Ed.)

[6] P. Christoffersen (2003). *Elements of Financial Risk Management.* Academic Press.

[7] R.D. Davé, G. Stahl (1998). On the Accuracy of VaR Estimates Based on the Variance-Covariance Approach. *Risk Measurement, Econometrics and Neural Networks*, 189-232.

[8] F.X. Diebold, R.S. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**, 253-263.

[9] 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.

[10] T. Gneiting (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106 (494)**, 746-762.

[11] A. McNeil, R. Frey, P. Embrechts (2015). *Quantitative Risk Management.* Princeton (2nd Ed.).

[12] C. A. G. Nass (1959). The 2-test for small expectations in contingency tables, with special reference to accidents and absenteeism. *Biometrika* **46**, 365-385.

[13] K. Pearson (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Mag. 5th Ser.* **50**, 157-175.

[14] D. Tasche (1999). Risk contributions and performance measurement. Working paper, Technische Universität München.

[15] D. Tasche (2002). Expected Shortfall and Beyond. *Journal of Banking & Finance* **26(7)**, 1519-1533.

[16] D. Tasche (2008). Capital allocation to business units and sub-portfolios: the Euler principle. In: Resti, A., editor, *Pillar II in the New Basel Accord: The Challenge of Economic Capital*, Risk Books, 423-453.