UPFBE
Working Paper Series

2021/2

Gyöngyi Bugár, Máté Uzsoki

Back-testing risk estimation models: a simulation study for two-asset portfolios

This publication was funded by the Project no. TKP2020-IKA-08, implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2020-4.1.1-TKP2020 funding scheme.

UNIVERSITY OF PÉCS
Faculty of Business and Economics

BACK-TESTING RISK-ESTIMATION MODELS: A SIMULATION STUDY FOR TWO-ASSET PORTFOLIOS

GYÖNGYI BUGÁR
Department of Management Science
Faculty of Business and Economics
University of Pécs
Pécs, Hungary
bugar.gyongyi@ktk.pte.hu

MÁTÉ UZSOKI
Serpentarius Software Ltd.
Budapest, Hungary
uzsoki.mate@serpentarius.hu

Abstract: The aim of the study is to check the validity of five different risk-estimation models for two-asset portfolios, a topic which is relevant in model selection both for determining the minimum capital requirements for trading book portfolios and for the regulatory monitoring of the performance of internal risk models. Simulations based on a real data set containing the FTSE 100 constituents were carried out, and the risk was gauged by Expected Shortfall, a measure which also captures tail risk. Given that the period studied includes that of the subprime crisis, there is an inherent opportunity to compare and contrast the results produced under disaster conditions with others from less stressful periods. Our empirical analysis has confirmed that using Expected Shortfall instead of Value-at-Risk alone is not enough, and that the risk model has to be carefully selected and back-tested. The general Pareto distribution proved to be a prudent choice for risk models. In fact, among the five models considered, the model when general Pareto marginals were coupled with Clayton copula showed the best performance. It was, however, also revealed that this model is susceptible to being "over-cautious" in estimating loss.

Keywords: Risk estimation models, Portfolio, Back-testing, Expected Shortfall, Copula

1 Introduction

Following the eruption of the subprime crisis in 2007 there has been a much greater need in the financial sector for valid, effective measures of risk, and this is of crucial importance in evaluating the performance of the trading book portfolios of banks. The recently published new Basel III standards (BCBS 2016) provide a revised framework for determining the capital charge for market risk in internal models with a shift from Value-at-Risk (VaR) to Expected Shortfall (ES), and this risk measure does improve the capture of tail risk. In spite of the fact that ES has gained a foothold in Basel III and is used as the proposed, new measure in estimating the risk of trading book positions, the Basel Committee still supports VaR for back-testing purposes.

However, there was a good deal of debate in the literature as to whether ES can be back-tested. The root of the debate was the lack of a certain property referred to as elicitability. Gneiting (2011) has pointed out that ES, in contrast to VaR, is not elicitable. It has led some voices to conclude that this risk measure is not back-testable at all (see for instance Carver 2013). Some others like Acerbi and Székely (2014) are convinced that this is not a real problem as they emphasize that elicitability has not even been exploited in the back-tests of VaR.

In this paper, we keep an open mind regarding the above-mentioned debate by taking an empirical standpoint. In fact, our aim is twofold. We first attempt to show how a particular back-testing procedure works in practice, whilst, in particular, trying to determine whether it reliably detects "bad performance" and whether it is capable of making a clear distinction between different risk-estimation models. Secondly, based on our back-testing results, we try to identify certain elements which may help us to devise better risk-estimation models.

The paper is structured as follows. In the next section a brief theoretical background to risk-estimation is provided, following which we describe the details of simulation modeling. We then present and discuss the results before closing the paper with some concluding remarks.

2 Theoretical Background

Estimating portfolio risk requires two important steps. The first one is the selection of an appropriate measure of risk, and the second is modeling statistical dependence, namely the co-movement between returns on different assets comprising the portfolio.

Markowitz (1952) was the first to explicitly involve risk as a decision parameter into the portfolio optimization process. He proposed making portfolio selection decisions relying on the expected return (E) and the variance of return (V). He used the former to gauge average profitability, and the latter to evaluate risk. For measuring dependence, i.e. the co-movement between the different pairs of asset returns, the linear (Pearson) correlation coefficients were used.

As Dowd (2005) concluded, in the case of real-world portfolio allocation decisions, two problems arise. The first is that widely used risk measures such as variance or VaR become unreliable (Günay (2017)). The second problem is that correlation-based (variance-covariance) approaches fail in correctly modeling statistical dependence. VaR can be determined as the highest possible loss at a certain confidence level for a given time interval. Based on the loss distribution function (*F*), at the confidence level α, VaR can be defined as the α-quantile of the loss distribution (Dowd and Blake 2006):

$$P\left(L \leq VaR_\alpha(L)\right) = F(VaR_\alpha(L)) = \alpha \tag{1}$$

where *P* indicates probability and *L* stands for loss. ES is defined as the expected value of losses exceeding VaR at a given confidence level (α) and time interval:

$$ES_\alpha(L) = E(L|L > VaR_\alpha(L)) \tag{2}$$

ES offers a possible solution to the first problem mentioned above, since it has more attractive theoretical and empirical properties than most of the other risk measures. First of all, it is a downside measure of risk and, hence, consistent with the intuitive notion of risk, since it takes into account only the unfavorable part of the return/loss distribution. Secondly, it is a coherent risk measure in the sense of the Artzner *et.al* (1999) axioms. Thirdly, it also accounts for losses beyond VaR, which is especially important in the case of fat-tail distributions. Finally, it has two favorable technical properties: it is continuous with respect to the confidence level and convex with respect to the control variables, the latter being highly relevant in portfolio optimization.[1]

[1] In order to optimize within the mean-ES framework, as it was shown by Rockafellar and Uryasev (2000), one has to solve a simple linear programming problem. It is worth mentioning that for ES they used the term Conditional Value-at-Risk (CVaR).

To the challenge of proper dependence modeling, as highlighted and illustrated by Dowd (2005), the application of copulas offers a tractable approach.

Back-testing or out-of-sample analysis[2] is a technique which makes it possible to evaluate the accuracy of a forecasting method relying on historical data (McNeil et al. (2015)). In the implementation two different non-overlapping time horizons are used. The first is the estimation period - which serves as a basis for estimating the intended financial variable (e.g. return or risk), and the second is the forecasting period - which can be used to compare its realized (real) value to that of the estimated one. In our analysis we intend to check the validity of different ES estimation models, and we do this by comparing the real portfolio loss with the estimated ES.

For back-testing purpose, we utilized the scoring function introduced by Acerbi and Székely (2014):

$$\mathrm{Z} = \frac{1}{(1-\propto)\mathrm{T}} \cdot \sum_{\mathrm{t}=1}^{\mathrm{T}} \frac{\mathrm{X_t I_t}}{\mathrm{ES}_{\propto,\mathrm{t}}} + 1 \qquad (3)$$

where $ES_{\alpha,t}$ is the estimated value of risk (measured by expected shortfall) in period *t* at a given confidence level α . X_t is the realized portfolio return in period *t,* and *T* is the total number of periods considered. *T* is the time horizon (normally a year) for which the evaluation is made. I_t is an indicator variable with the value of 0 or 1. It is 0 when $X_t + VaR_{\alpha,t} \geq 0$ and 1 when $X_t + VaR_{\alpha,t} < 0$. In the latter case $-X_t = L_t > VaR_{\alpha,t}$, i.e. the realized loss is higher than the estimated value of $VaR_{\alpha,t}$. We refer to this case as a violation. In an ideal case when the model predicts ES perfectly, the Z-score is equal to zero. A positive value of Z indicates an overestimation of ES, whilst a negative value means that the risk is underestimated.

3 Simulation Modeling

3.1 Portfolio Models

The marginal return distributions of the portfolio components were modeled either using normal or generalized Pareto distribution (GPD). As is well-known, the normal (or Gaussian) distribution can be uniquely described with two parameters - the mean and the standard deviation. The GPD can be specified by three parameters - the first for the location (μ), the second for the scale (σ) and the third for the shape (ξ).[3] This type of distribution has proved to be useful in modeling tail risk. The dependence structure was modeled relying on various copula models, and, in particular, as a benchmark case, we used the linear correlation coefficient or the so-called Gaussian copula. In addition, there were two other one-parameter Archimedean copulas fitted to the data - a Clayton and a Gumbel copula. The reason for our choice is that these types of copula have proved to be useful in modeling the simultaneous occurrence of major losses.

---

[2] It is also known as ex ante analysis.

[3] The GPD can be characterized by the cumulative distribution function as follows (Coles (2001)).

$$F_{(\mu,\sigma,\xi)}(x) = \begin{cases} 1 - \left(1 + \dfrac{\xi(x-\mu)}{\sigma}\right)^{-\frac{1}{\xi}} & \text{for } \xi \neq 0 \\ 1 - \exp\left(-\dfrac{x-\mu}{\sigma}\right) & \text{for } \xi = 0 \end{cases}$$

for $x \geq \mu$ when $\xi \geq 0$, and $\mu \leq x \leq \mu - \frac{\sigma}{\xi}$ when $\xi < 0$,

where μ, σ, ξ are real numbers, and $\sigma > 0$.

The bivariate Clayton and Gumbel copulas can be defined, respectively, as follows (Nelsen (2006)):

$$C_\theta(v_1, v_2) = max\left\{\left[{v_1}^{-\theta} + {v_2}^{-\theta} - 1\right]^{-\frac{1}{\theta}}, 0\right\} \text{ for } \theta \geq -1,\ \theta \neq 0 \quad (4)$$

$$C_\theta(v_1, v_2) = exp\left\{-\left[(-lnv_1)^\theta + (-lnv_2)^\theta\right]^{\frac{1}{\theta}}\right\} \text{ for } \theta \geq 1 \quad (5)$$

where v_1 and v_2 are standard uniform random variables (i.e. they are distributed evenly over the range [0,1]).

Considering all of the possible cases – except one[4] – with the two different marginals and the three dependence structures, all in all we have studied five different models. They are presented in Table 1.

Table 1 Portfolio models simulated

Name of Model	Marginals	Copula
Normal-Gaussian	Normal	Gaussian / Linear correlation
Normal-Clayton	Normal	Clayton
Normal-Gumbel	Normal	Gumbel
Pareto-Clayton	GPD	Clayton
Pareto-Gumbel	GPD	Gumbel

Source: own work

As a profitability measure, the expected return (mean) was considered, and the risk was measured by ES.

3.2 Simulation Process

Before performing the Monte Carlo simulation, parameter estimations were carried out. For each out-of-sample period we estimated the parameters of the normal and the Pareto distribution from real market data. Also, the copula parameters were estimated for each pair of equities to describe the co-movements between their returns.

Following the estimation of the necessary parameters, we generated returns for the different portfolio models by Monte Carlo simulation. First, a two-variable standard uniform distribution with the intended copula was simulated, and then the intended marginals were fitted on the given copula. The steps of this process can be described as follows (see also Bouyé *et al.* 2000; Dowd 2005).

- Generate two independent random variables from a standard uniform distribution: v_1, v_2.

- Set $u_1 = v_1$. (6)

[4] In fact, we have not considered GPD marginals coupled with Gaussian copula.

$$C(u_2|u_1) = \frac{\partial C(u_1,u_2)}{\partial u_1} = v_2. \quad (7)$$

- Solve (7) for u_2.
- Use the classical inversion method to obtain random values for the intended marginal:

$$F^{-1}(u_i) = r_i \qquad (i = 1,2). \quad (8)$$

- Obtain a simulated portfolio return with a portfolio-weight *w*:

$$R = wr_1 + (1 - w)r_2. \quad (9)$$

- Repeat the steps *m* times (in our case $m = 10000$).
- Determine ES based on the simulated return distribution.

4 Results

4.1 Data and Descriptive Statistics of Individual Equities

In the empirical analysis, 25 two-asset portfolios were studied by randomly selecting various couples of company shares from among the FTSE 100 constituents. The estimation of the marginals and that of the dependency structure of the return distributions was based on the above-mentioned real-data set comprising the daily closing prices of the relevant equities for the time interval of 16 years, stretching from January 4, 2000 until December 31, 2015. For the equities involved in the analysis, daily percentage returns were calculated.

There were, all in all, 31 individual equities included in the 25 two-asset portfolios considered. Table 2 shows the descriptive statistics of the daily returns calculated for the three adjoining five-year periods from January 2, 2001 until December 31, 2015. All the average values – except for the correlation presented in the last column – given in the table must be interpreted as grand means - namely as the results of averaging done over time and equity also.[5] The average correlation is calculated over time, and the 25 equity pairs comprising the portfolios are considered. The average, minimum and maximum returns are presented as percentages.

Table 2 Descriptive statistics of the 31 individual equities included in the two-asset portfolios for the three adjoining five-year periods between 2001 and 2015

Period	Average Return (percent)	Minimum Return (percent)	Maximum Return (percent)	Average SD	Average Skewness	Average Kurtosis	Average Correlation
2001-2005	0.028	-62.97	23.89	0.020	0.113	10.15	0.248
2006-2010	0.034	-66.57	73.24	0.023	-0.192	13.18	0.381
2011-2015	0.029	-22.10	18.43	0.015	0.323	12.43	0.377

Source: own calculations based on daily closing prices of FTSE 100 constituents provided by Bloomberg

It is not surprising that the most extreme return, standard deviation, kurtosis and correlation values appeared in the second period – which covered the subprime crisis. Indeed, between 2006 and 2010 the average daily return was 3.4 basis points, with minimum and maximum

[5] Because of the high number of equities, we do not report individual values.

returns of -66.57 and 73.24 percent, respectively. The average standard deviation, kurtosis and correlation values were 0.023 (2.3 percent), 13.18 and 0.381, respectively. It is also notable that the average skewness was negative (-0.192) in the same period. The normality of the daily return was tested for all equites in each period, and none of the return distributions proved to be normal at 5 percent significance level.

4.2 Simulation Results

There was a sliding window of 250 days (in practical terms, a year) used as an estimation period. After having estimated the ES for the day following the first year, we shifted the in-the-sample estimation period one day forward and made a new risk-estimation for the subsequent day. In total, with this rolling technique, we generated 3750 non-overlapping out-of-sample ES estimations for each portfolio models[6]. In accordance with the recent Basel III regulation (BCBS 2016) ES has been estimated as having a 97.5 percent confidence level.

Based on the estimated daily ES values, the Z-score given in (3) was determined for each of the 25 portfolios considered. We studied equally weighted portfolios, with 50-50 percent invested in each of the two equities. The time horizon (T) in calculating Z was set to be 250. It made possible to compare and contrast the performance of the different models on a yearly basis.

Figure 1 shows the average Z-scores for the 25 randomly selected portfolios given by the different portfolio models over the forecasting period from 2001 until 2015. It can be concluded that all models considered normally agreed in either over- or underestimation the capital charge for risk. Indeed, there were only two years (2010 and 2013) when two of the models, and in particular those with the Clayton dependency structure, on average overestimated the ES value, whilst the other three underestimated it. Further, there is a periodic change in under- and overestimation. Among the five models, the Pareto-Clayton showed the best performance, but, in fact, in 7 of the 15 years it produced the lowest average Z-score (in absolute value). It is notable that it proved to be the most reliable in those years when faced with risk underestimation - which produced its heaviest effects in 2007 and 2008. The Normal-Clayton model emerged as second best, also showing a typically good performance in those years when there was a general tendency to underestimate risk. However, both of these models proved to be “over-cautious” in comparison to the others in the years when the risk was overestimated. The Pareto-Gumbel model showed roughly the same performance as the Normal-Clayton with the distinctive feature of being more successful than the latter in cases of overestimation and less successful when the risk was underestimated. In the rank order of models Normal-Gumbel can be placed last following Normal-Gaussian.[7]

Figure 2 presents the average daily estimated value of ES given by the different portfolio models for the time horizon of 2001-2015. It should be noted that the discrete points belonging to the particular years are connected with straight lines. It is clearly visible that, for the models with normal marginals, ES seems to be more stable over time, and moves on a less high level than for the two models with Pareto marginals. In 2002, for instance, the average daily ES with the value indicating almost 30 percent loss for the Pareto models was about six times higher than that of the approximately 5 percent estimated for the models with normal marginals.

[6] We have built a C#.Net application using R Statistics and Azure SQL to run the simulations.

[7] If we need the ranking order of the different models, we should rank them in each year based on the absolute value of their average Z-score a (lower number indicates a better performance) and summarize the yearly rank numbers. As aggregate rank numbers, this produces 37 for the Pareto-Clayton, 44 for the Pareto-Gumbel and Normal-Clayton, 47 for the Normal-Gaussian and 53 for the Normal-Gumbel models, respectively.

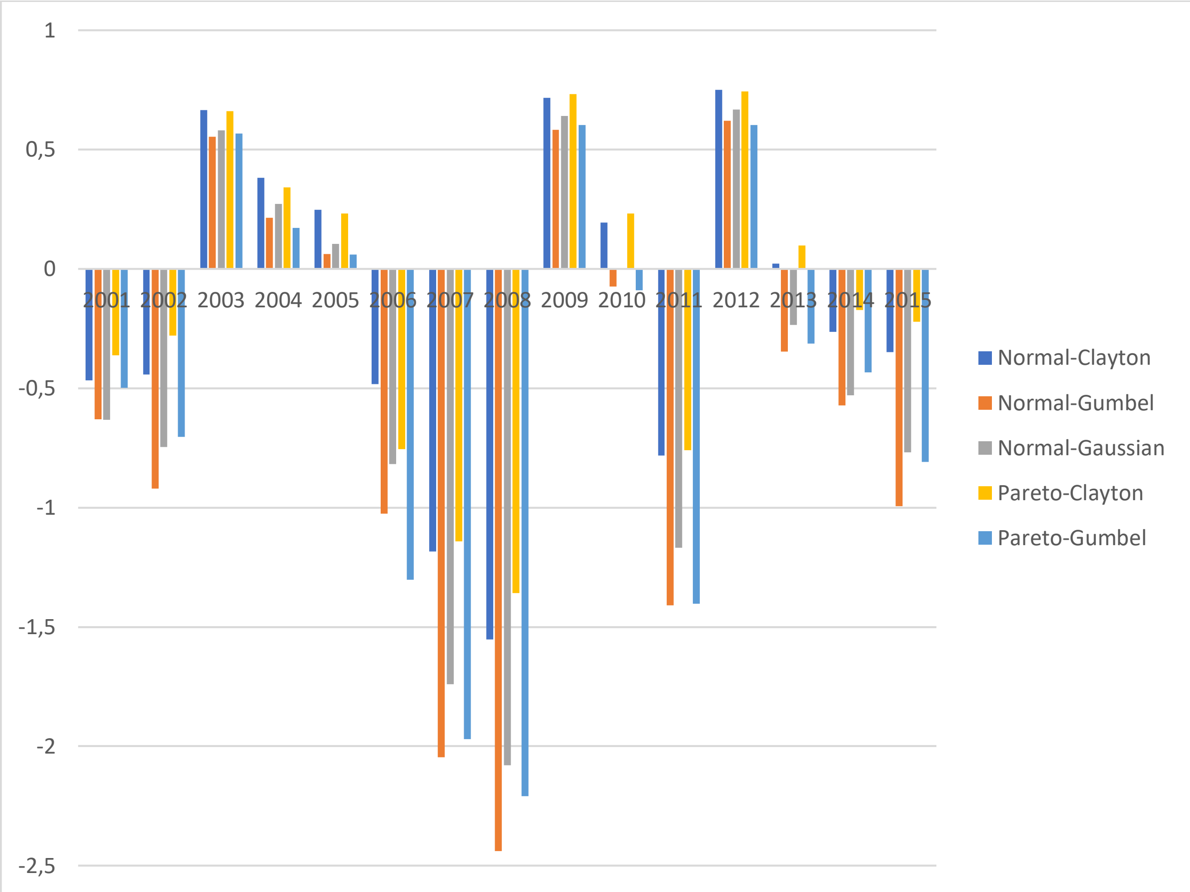

Figure 1 Average Z-score of 25 two-asset portfolios given by the different models between 2001 and 2015

Source: own work

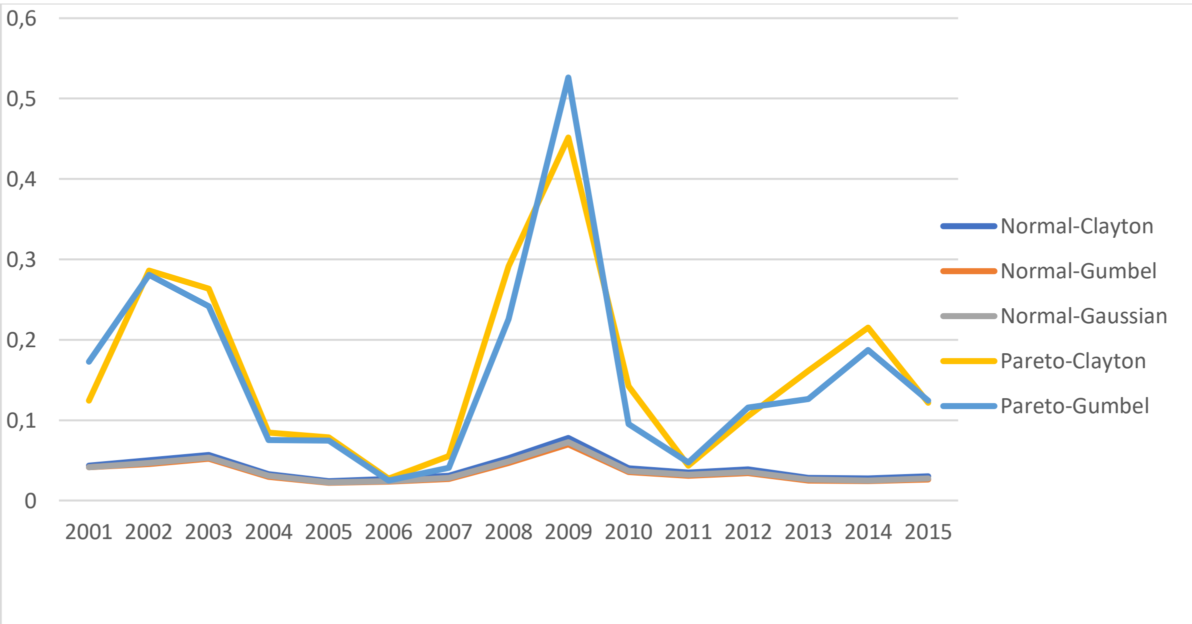

Figure 2 Average daily estimated value of Expected Shortfall (ES) for the different portfolio models between 2001 and 2015

Source: own work

This can, however, be justified (see Figure 1) by the superior performance of models with Pareto marginals over those with normal. During the financial crisis - in 2008 - we can see a similar pattern, although in 2009 the still higher estimated value of Expected Shortfall for Pareto models did not prove to be necessary to cover real losses. Indeed, as can be seen in Figure 1, each model overestimated the risk for this year, and the extent of overestimation proved to be higher in the case of Pareto models. We also examined how the different copula models coupled with the two given marginals performed. Figures 3 and 4 show the average Z-score of the 25 two-asset portfolios for the copula models studied when the marginals were fixed to be normal and GPD, respectively.

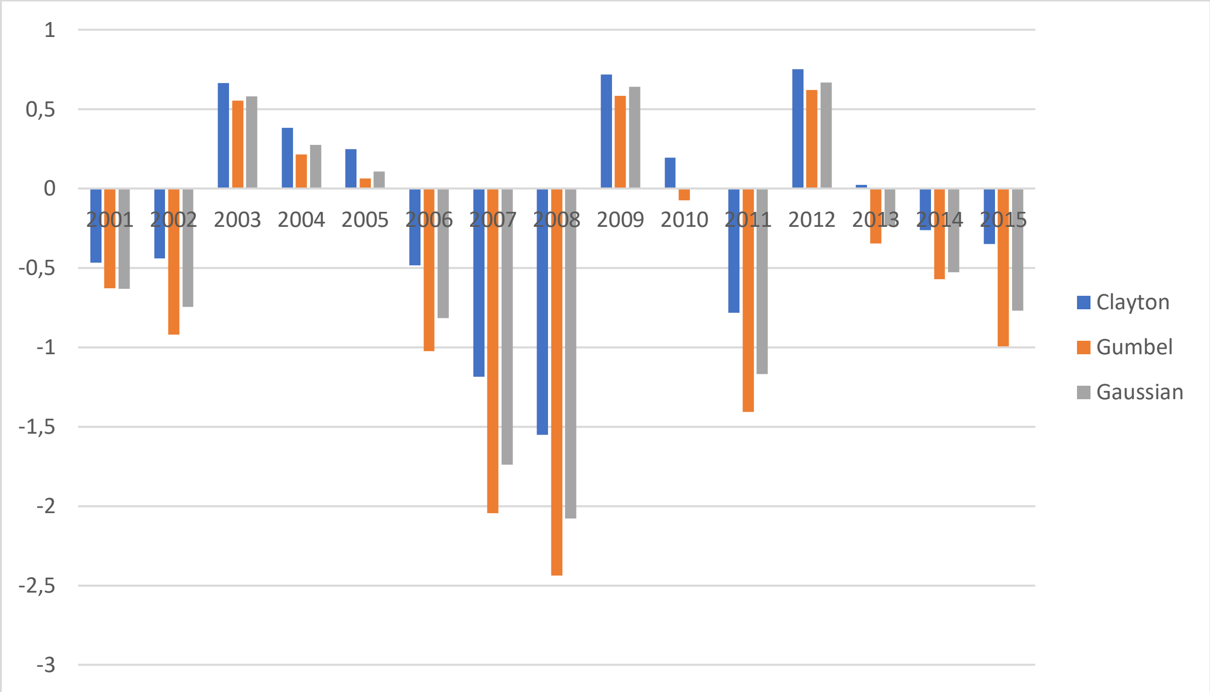

Figure 3 Average Z-score of 25 two-asset portfolios given by different copula models with normal marginals between 2001 and 2015

Source: own work

Figure 3 shows that, relying on normal marginals, the models using Clayton copula in general[8] outperformed those featuring Gumbel and Gaussian dependency structures. In particular, in 9 cases out of 15, especially when the real losses were underestimated the average Z-score was the lowest for models with Clayton copula. Precisely the opposite applies to those years (e.g. 2003 or 2009) when we experienced overestimation. In the latter cases the Gumbel copula proved to be the best.

Relying on GDP marginals (see Figure 4) the overall superior performance of models with Clayton to those with Gumbel copula can also be confirmed. In fact, the Clayton models show better performance in 9 years, and – similarly to the models with normal marginals – this occurred in the cases of risk-underestimation (with one exception).

[8] The aggregate rank numbers for Clayton, Gaussian and Gumbel dependency structures are 27, 30, 37, respectively.

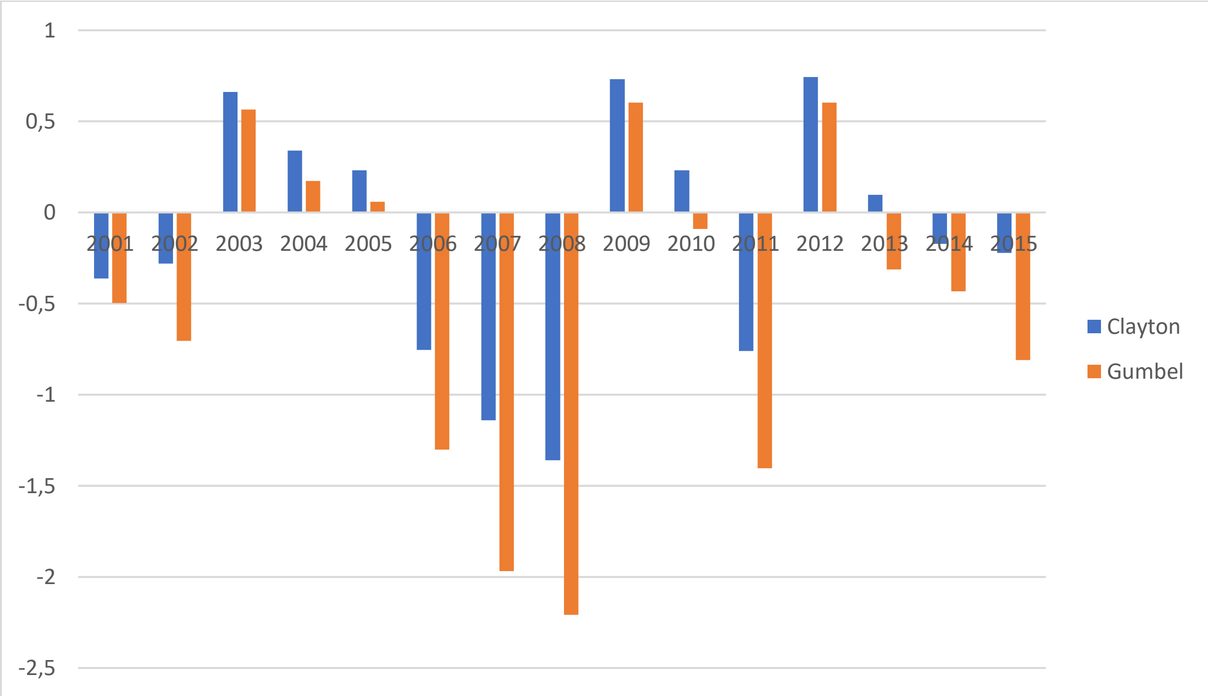

Figure 4 Average Z-score of 25 two-asset portfolios given by different copula models with GDP marginals between 2001 and 2015

Source: own work

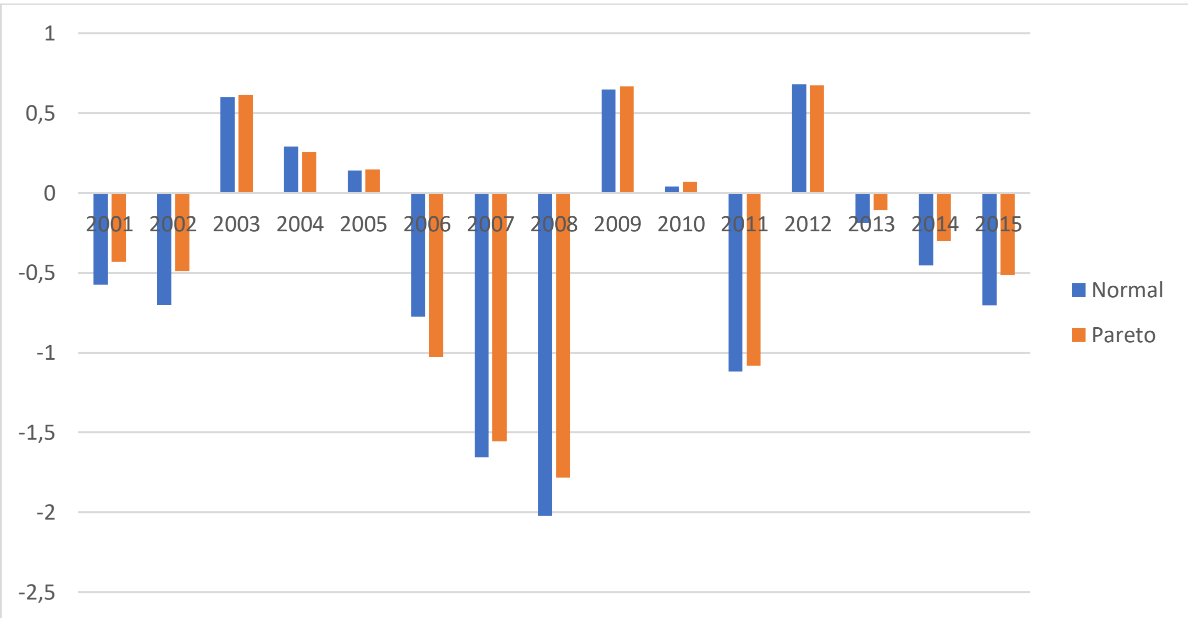

Figure 5 Average Z-score of 25 two-asset portfolios given by different marginals between 2001 and 2015

Source: own work

In addition, the effect of choosing marginals for forecasting performance was studied, and Figure 5 shows the results. The superior performance of GPD models is remarkable. Indeed, in twice as many cases (10 as opposed to 5) the models with GPD marginals performed better than their counterparts with normal marginals.

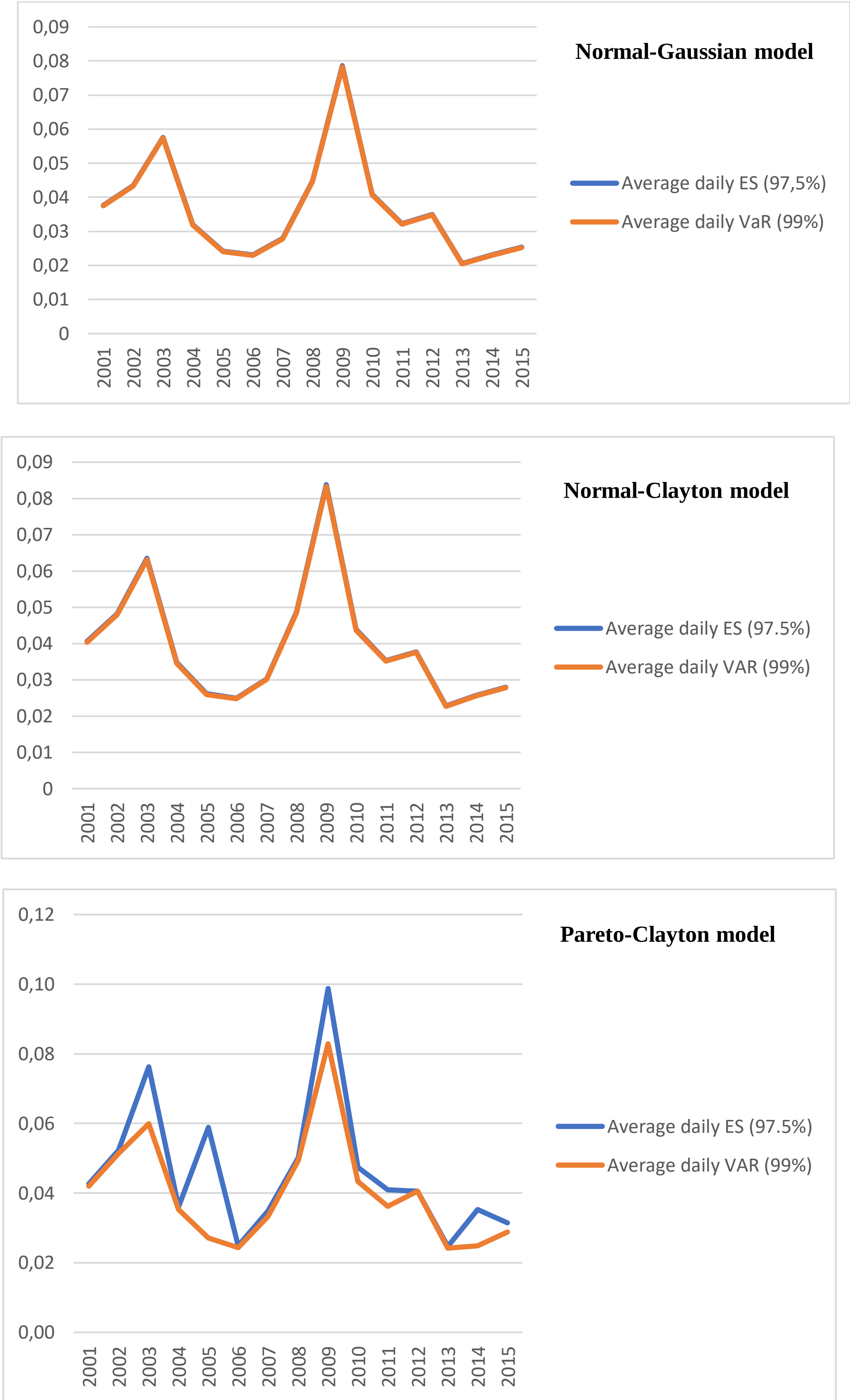

Figure 6 Average daily estimated ES and VaR of a randomly selected two-asset portfolio between 2001 and 2015

Source: own work

Finally, we try to illustrate the significance of model selection on the estimated risk given by both the old and the new capital requirements regime. Figure 6 shows the average daily VaR and ES of a randomly selected two-asset portfolio estimated at 99 and 97.5 percent confidence levels, respectively.[9] The discrete points belonging to the specific years are connected with straight lines.

Whilst the profile of the VaR curve is almost identical in all three cases, the values are slightly higher when a Clayton copula was fitted to normal marginals, and even higher when Pareto marginals were coupled with Clayton copulas. In addition, Figure 6 shows that, for the Normal-Gaussian model, the ES is nearly equal to its VaR counterpart. This is no surprise since, for such a benchmark distribution, the 97.5 percent confidence level for ES has been calibrated for the purpose of producing the same capital charge as the old 99 percent-VaR regime. It is also notable that changing the dependency structure (from the Gaussian copula to the Clayton) has not induced any change in this feature. Indeed, it is also true for the Normal-Clayton model that the 97.5 percent-ES and the 99 percent-VaR curves practically coincide - which is in sharp contrast to our findings for the Pareto-Clayton model. In this case, the new regime provided higher risk estimates than the old.

5 Concluding remarks

For this paper, we performed back-tests on five different risk-estimation models for 25 equally weighted, two-asset portfolios. These portfolios were constructed by randomly selecting various couples from among the FTSE 100 constituents. For modeling the marginals, the normal and general Pareto distributions, and, with respect to the dependency structure, the Gaussian, the Clayton and Gumbel copulas were fitted to the data and simulated. Compared to the benchmark model (i.e. normal marginals with the Gaussian copula) using Pareto marginals makes it possible to cope better with tail risk, whilst daily portfolio returns for each model were generated by the Monte Carlo simulation after the necessary parameters had been estimated. By relying on 10000 simulated returns, daily ES values were estimated, and realized returns (losses) were compared to them. For back-testing purposes we utilized a scoring function proposed by Acerbi and Székely (2014).

Our empirical analysis has confirmed that using ES instead of VaR alone is not enough, and that the risk model has to be carefully selected and back-tested. The general Pareto distribution proved to be a prudent choice for risk models. In fact, among the five models considered, the model when GPD marginals were coupled with Clayton copula showed the best performance. It was, however, also clear that this model is susceptible to being “over-cautious” in estimating loss. For two-asset portfolios the choice of marginals seems to be more significant than that of copula models – a finding which is in line with that of Low *et al.* (2013).

There are some limitations of our current study that open a new perspective for additional research. First of all, the time period utilized for empirical analysis can be extended. In particular, involving the crisis-period of the pandemic can give a further insight into the performance of risk estimation models under stressful conditions. The effect of using better suited copulas could be different in multi-asset portfolios, and this also offers an interesting direction for future investigation. Additionally, we might propose and test dynamic or adaptive models in an attempt to overcome the problem that prudent risk measures tend to produce higher risk estimates in stable periods also.

Acknowledgement: Project no. TKP2020-IKA-08 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2020-4.1.1-TKP2020 funding scheme.

[9] Both FTSE constituents, not named here, were represented by 50 percent weight in the portfolio.

References

Acerbi, C., Székely, B., 2014. Backtesting Expected Shortfall. *Risk Magazine*, 27, pp. 76-81.

Artzner, P., Delbaen, F., Eber, J., Heath, D., 1999. Coherent Measures of Risk. *Mathematical Finance*, 9, pp. 203-228.

Basel Committee on Banking Supervision, 2016. Minimum capital requirements for market risk. Bank for International Settlements. [online] Available at: http://www.bis.org/bcbs/publ/d352.pdf (Accessed 31 March 2017).

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., Roncalli, T., 2000. Copulas for Finance – A Reading Guide and Some Applications. Groupe de Recherce Opérationelle, Crédit Lyonnais, Paris. [online] Available at: http://66ssrn.com/abstract=1032533 (Accessed 18 August 2015).

Carver, L., 2013. Mooted VAR substitute cannot be back-tested, says top quant. *Risk,* 8, March. [online] Available at: https://www.risk.net/regulation/baselcommittee/2253463/mooted-var-substitute-cannot-be-back-tested-says-top-quant (Accessed 20 April 2018)

Coles, S., 2001. *An Introduction to Statistical Modeling of Extreme Values*, Springer London Ltd, London.

Dowd, K., 2005. Copulas and Coherence – Portfolio Analysis in a Non-Normal World. *Journal of Portfolio Management*, Fall, pp. 123-127.

Dowd, K., Blake, D., 2006. After VaR: The Theory, Estimation, and Insurance Applications of Quantile-Based Risk Measures. *Journal of Risk and Insurance*, 73(2), pp. 193–229.

Gneiting, T., 2011. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106, pp. 746-762.

Günay, S., 2017. Value at risk (VaR) analysis for fat tails and long memory in returns. *Eurasian Economic Review,* 7, pp. 215–230.

Low, R. K. J., Alcock, J., Faff, R., Brailsford, T., 2013. Canonical vine copulas in the context of modern portfolio management: Are they worth it? *Journal of Banking and Finance*, 37, pp. 3085-3099.

Markowitz, H. M., 1952. Portfolio Selection. *Journal of Finance*, 8, pp. 77-91.

McNeil, A. J., Frey, R., Embrechts, P., 2015. *Quantitative Risk Management – Concepts, Techniques and Tools*. Princeton University Press, Princeton, Oxford.

Nelsen, R. B., 2006. *An Introduction to Copulas*. Springer-Verlag, New York.

Rockafellar, R. T., Uryasev, S., 2000. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3), pp. 21-41.