Risk Measures Prediction and Its Sensitivity to the Refit Step: A Score-Driven Approach

Kateřina Nováková[1], Petra Tomanová[2]

Abstract. The aim of this paper is an assessment of the refit step impact on financial risk measures prediction – Value at Risk (VaR) and Expected Shortfall (ES) – for four world market price indices. Generalized Autoregressive Score (GAS) models assuming the Student's *t*, skew-Student's *t*, and Gaussian distributions are analyzed and compared to the t-GARCH model. VaR and ES predictions are backtested using rolling windows while considering various refit steps. Three different performance measures for predictions are utilized: dynamic quantile test, quantile loss function, and Fissler and Ziegel loss function. The results show that the choice of the refit step does not significantly influence VaR and ES predictions based on GAS models with the Student's *t* and skew-Student's *t* distribution. However, VaR and ES predictions based on Gaussian distribution react extensively in the periods of price shocks.

Keywords: expected shortfall, generalized autoregressive score model, prediction, value at risk

JEL Classification: C22
AMS Classification: 91G70

1 Introduction

Risk measures evaluate the risks that a financial institution goes through. The two leading risk measures are Value at Risk (VaR) and Expected Shortfall (ES). The VaR measures the largest expected portfolio loss over a particular time horizon at a given probability level assuming normal market conditions. It can also be comprehended as an estimate of the largest loss that could occur with $100\alpha\%$ probability based on already known losses within a certain period of time. Despite being widely used by all banks and regulators, VaR does not fulfill one of the axioms of coherence [1]. These axioms strive to distinguish *good* and *bad* risk measures. Breaking some of them can lead to paradoxical results. Specifically, the VaR violates subadditivity since a sum of portfolios might exhibit a higher risk (VaR) than sub-portfolios. The ES is introduced by Rockafellar et al. [6] and it measures the expected loss in the $100\alpha\ \%$ worst cases where usually $\alpha \in \{0.01, 0.05\}$ and therefore it takes into consideration the shape of the distribution tail. However, it does not consider only the worst case that can occur but the average of the worst cases. The ES is proven to be a coherent indicator [5].

The risk measure estimation requires an accurate estimate of the conditional distribution of future returns. Then, the VaR and ES at time *t* for a risk level α can be computed as

$$VaR_t(\alpha) \equiv F^{-1}(\alpha; \boldsymbol{\theta}_t, \boldsymbol{\xi}), \qquad ES_t(\alpha) \equiv \frac{1}{\alpha}\int_{-\infty}^{VaR_t} z dF(z, \boldsymbol{\theta}_t, \boldsymbol{\xi}),$$

where F^{-1} denotes the inverse of the continuous cumulative density function, $\boldsymbol{\theta}_t$ is a vector of time-varying parameters and $\boldsymbol{\xi}$ is a vector of additional static parameters. Thus, the VaR is simply the $100\alpha\ \%$ quantile of the return distribution at time *t* and the ES is the average of the $100\alpha\ \%$ worst cases.

In this paper we utilize the Generalized Autoregressive Score (GAS) framework for the time-varying parameter estimation. First, the GAS models are defined in Section 2. Second, the GAS estimates are compared to the well-known GARCH models on an empirical study of four stock market indices in Section 3. Third, in Section 4, the VaR and ES estimates are backtested using rolling windows and the impact of the length of the refit step is analyzed. Section 5 concludes.

[1] University of Economics, Prague, Department of Econometrics, W. Churchill Sq. 1938/4, 130 67 Prague 3, Czech Republic, katerina.novakova1@seznam.cz
[2] University of Economics, Prague, Department of Econometrics, W. Churchill Sq. 1938/4, 130 67 Prague 3, Czech Republic, petra.tomanova@vse.cz

2 GAS Models

Generalized autoregressive score (GAS) models proposed by Creal et al. [4] belong to the class of observation-driven models that utilize a scaled score of the likelihood function as the driving mechanism. The possibility to let some parameters vary in time is necessary for capturing the dynamic behavior of time series. A huge benefit of GAS models is their ability to take advantage of the complex density structure rather than only consider means and higher moments. Moreover, the likelihood evaluation is straightforward.

In accordance with a notation of [3], let $\boldsymbol{y}_t$ be an N-dimensional random vector of the dependent variables at time t and $\boldsymbol{\theta}_t$ be a vector of time-varying parameters. Then $\boldsymbol{y}_t$ follows conditional observation density $p(\cdot)$

$$\boldsymbol{y}_t|\boldsymbol{y}_{1:t-1} \sim p(\boldsymbol{y}_t; \boldsymbol{\theta}_t)$$

for $t = 1, \ldots, T$, where $\boldsymbol{y}_{1:t-1}$ is a matrix which contains the past values of $\boldsymbol{y}_t$ up to time $t-1$. The vector of time-varying parameters $\boldsymbol{\theta}_t$ depends on $\boldsymbol{y}_{1:t-1}$ and a set of additional static parameters $\boldsymbol{\xi}$, $\boldsymbol{\theta}_t \equiv \boldsymbol{\theta}(\boldsymbol{y}_{1:t-1}, \boldsymbol{\xi})$.

The GAS updating mechanism for the time-varying parameter $\boldsymbol{\theta}_t$ is

$$\boldsymbol{\theta}_{t+1} \equiv \boldsymbol{\kappa} + \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{B}\boldsymbol{\theta}_t,$$

where $\boldsymbol{\kappa}$ is a vector of constants measuring the level of the process, $\boldsymbol{B}$ is a diagonal matrix of autoregressive coefficients controlling for the persistence of the process and $\boldsymbol{A}$ is a diagonal matrix of parameters indicating the step of the update. $\boldsymbol{\kappa}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are collected in the set $\boldsymbol{\xi}$. $\boldsymbol{s}_t$ is the scaled score, which depends on the past observations and the time-varying parameters

$$\boldsymbol{s}_t \equiv \boldsymbol{S}_t(\boldsymbol{\theta}_t)\boldsymbol{\nabla}_t(\boldsymbol{y}_t, \boldsymbol{\theta}_t),$$

where $\boldsymbol{S}_t$ is the scaling function and $\boldsymbol{\nabla}_t$ is the score

$$\boldsymbol{\nabla}_t(\boldsymbol{y}_t, \boldsymbol{\theta}_t) \equiv \frac{\partial \log p(\boldsymbol{y}_t; \boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t}, \qquad \boldsymbol{S}_t(\boldsymbol{\theta}_t) \equiv \mathcal{I}_t(\boldsymbol{\theta}_t)^{-\gamma}.$$

Creal et al. [4] suggest to set the scaling matrix to the γ-th power of the Fisher information matrix

$$\mathcal{I}_t(\boldsymbol{\theta}_t) \equiv \mathsf{E}_{t-1}\Big[\boldsymbol{\nabla}_t(\boldsymbol{y}_t, \boldsymbol{\theta}_t)\boldsymbol{\nabla}_t(\boldsymbol{y}_t, \boldsymbol{\theta}_t)'\Big].$$

The vector of static parameters $\boldsymbol{\xi} \equiv (\boldsymbol{\kappa}, \boldsymbol{A}, \boldsymbol{B})$ can be estimated by maximizing the log-likelihood function.

3 Empirical Study

Four major world stock market indices are analyzed in the empirical study. The first two indices, DJIA and S&P 500, are related to the U.S. stock market. The FTSE 100 assesses the market in Great Britain and TOPIX covers the Japanese market. Two time periods are analyzed: (i) January 3, 2000 – December 31, 2010, which covers 2,767 days, (ii) January 4, 2010 – March 15, 2019, which covers 2,315 days. The first period contains the global financial crisis and the second represents recent years. Each of the chosen periods evinces diverse shapes of the return distribution which allows to overview of each model and distribution reaction. The dataset is downloaded from Thomson Reuters Datastream.

3.1 Comparison of GAS Models

GAS models can utilize a wide range of possible conditional distributions. However, since the price returns are often fat-tailed and possibly skewed, the most common distributions are Student's t and skew-Student's t. This property is often demonstrated by comparing the results with the Gaussian distribution which is symmetric and very sensitive to extreme values and changes in return variance. The scale parameter is treated as time-varying which follows the properties of price returns. The skewness and kurtosis parameters of the related distribution are tested whether the parameters vary over time. Dynamics for both of them are not statistically significant across various time periods, thus skewness and kurtosis are treated as constant and the scale is the only dynamic parameter. Since different scaling functions have no significant effect on results, it is set to identity.

GAS models are estimated for each price index over each time period and evaluated based on the Akaike information criterion (AIC). The results for both periods are shown in Table 1. The values of AIC are the

	2000–2010				2010–2019			
AIC	**DJIA**	**S&P 500**	**FTSE 100**	**TOPIX**	**DJIA**	**S&P 500**	**FTSE 100**	**TOPIX**
Student's t	7.966	8.275	8.177	8.962	5.267	5.457	5.819	6.825
Skew-Student's t	7.968	8.276	8.177	8.966	5.269	5.459	5.820	6.826
Gaussian	8.073	8.354	8.199	9.005	5.418	5.629	5.885	6.948

Table 1 AIC values for GAS models in 2000–2010 and 2010–2019

lowest for the model utilizing the conditional Student's t distribution, however, the differences between its skewed version are negligible. Models based on normal distribution perform a lot worse as expected.

Figure 1 compares individually estimated ES series for GAS models with Student's t and Gaussian distributions for two indices S&P 500 and TOPIX in the periods of 2000–2010 and 2010–2019. The green and blue lines correspond to the Gaussian and the Student's t distribution respectively. Both periods are characterized by a different behavior. While the first one (2000–2010) exhibits higher returns fluctuations due to the financial crisis, the second one (2010–2019) is more tranquil with occasional jumps. These properties result in different estimates of risk measures as well as the sensitivity of the index itself.

The S&P 500 fluctuates less and therefore does not exhibit too many sudden drops. On the other hand, TOPIX is more sensitive and occasional jumps result in significant drops in estimated risk measures. In the period of 2010–2019, the drops are more severe, e.g. the estimated ES of TOPIX drops to -20.158 in 2011. It confirms that the Gaussian distribution results in undue sensitivity to extreme values since it does not treat the return heavy-tails properly.

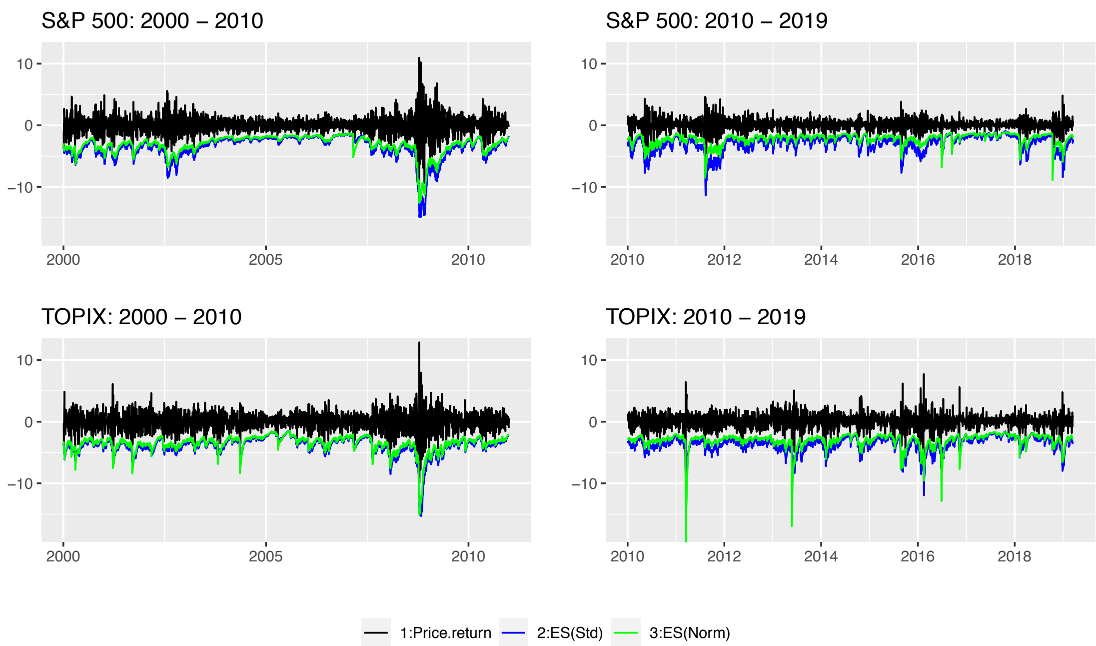

Figure 1 The estimated ES for S&P 500 and TOPIX based on Student's t and Gaussian distributions

3.2 Comparison of GAS and GARCH Models

GARCH models are so far one of the most often applied volatility models. Both GARCH and GAS models belong to the class of observation-driven models. Moreover, the original GARCH model is a special case of the GAS model, specifically, they coincide when the normal distribution and the inverse of Fisher information are utilized. However, it does not apply to the t distribution since t-GARCH and t-GAS updating mechanisms differ. Generally, GAS models have the advantage of using the complex density structure rather than only means and higher moments.

We compare t-GARCH and t-GAS on an example of TOPIX and S&P 500 in the periods of 2000–2010 and 2010–2019. Results in Table 2 show that while the t-GARCH performs better for S&P 500 in terms of the AIC, it is the opposite case for the more volatile TOPIX where the t-GAS is superior. The estimated VaR alongside the estimated volatility is plotted in Figure 2 where the blue and purple lines represent the estimated volatility and VaR from t-GARCH model respectively, the red and green lines refer to the estimated volatility and VaR from t-GAS model respectively, and the returns are black colored. For the TOPIX, the differences in estimated VaR are rather small. However, VaRs for S&P 500 exhibit noticeable differences. As expected, the score of the t distribution in GAS dynamics avoids the volatility to react too extensively to large values of returns. The idea is that such large values might be caused by the fat-tailed nature of the data and thus, should not be fully attributed to increases in the variance.

	2000–2010		**2010–2019**	
AIC	**t-GARCH**	**t-GAS**	**t-GARCH**	**t-GAS**
S&P 500	8.266	8.275	5.440	5.457
TOPIX	8.963	8.962	6.833	6.825

Table 2 The AIC values for t-GARCH and t-GAS

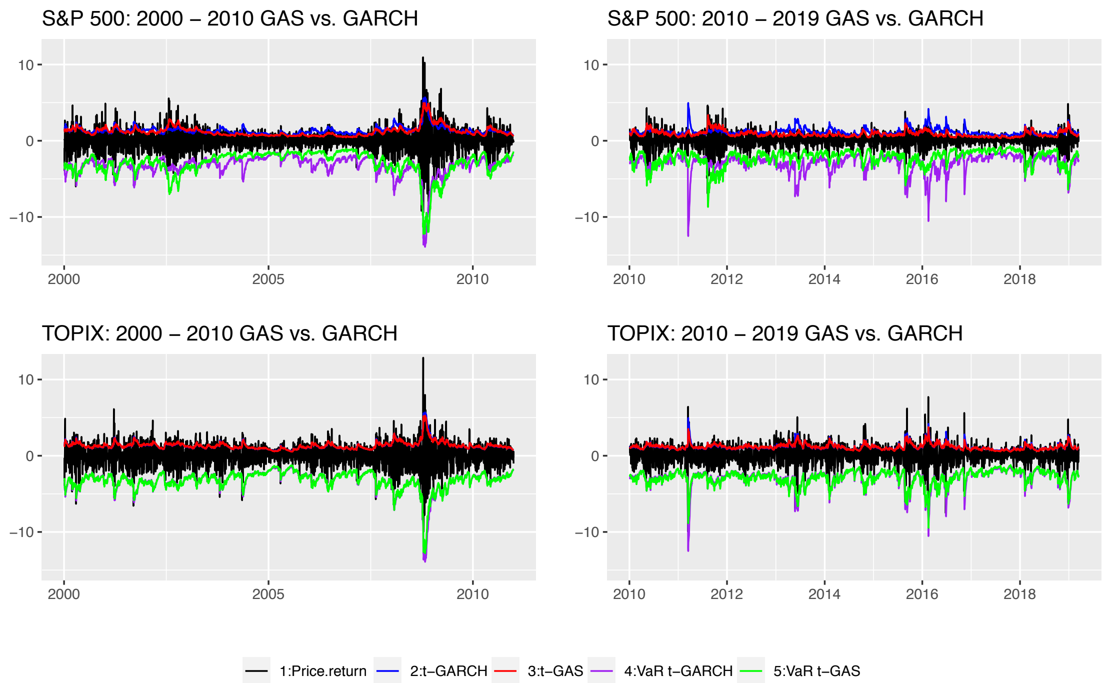

Figure 2 The estimated 1% VaR for S&P 500 and TOPIX based on t-GAS and t-GARCH models

4 Backtesting VaR and ES for GAS Models

Backtesting verifies the precision of VaR and ES predictions. The sample of length T is divided into two parts: in-sample of the length m and out-of-sample of the length $T-m$, and the approach of rolling windows is utilized. We analyze different refit steps for rolling windows and their impact on predictions. The output of the rolling windows are predicted values of the length $T-m$ and they are used to calculate VaR and ES. Then, the models are assessed by the dynamic quantile (DQ) test and the Fissler and Ziegel loss (FZL) function for the joint VaR and ES evaluation.

4.1 Sensitivity of the Refit Step for Rolling Windows

There are two parameters that in rolling windows that are required to be set: the forecast length $T - m$ and the length of the refit step. The forecast length is set to 1000 which represents approximately one-third of the whole sample. The length of the refit step can vary from 1 to T and there is no rule how to set it. The natural choice seems to be the length of 1 and 5 for the daily data [2] and the length of 4 for quarterly data [3]. Therefore, the sensitivity of the refit step is analyzed for the daily returns by comparing the estimated VaR and ES based on GAS models. Considered lengths are 1, 5, and 30.

For the GAS model utilizing the Student's *t* distribution, the lengths of 1 and 5 result in almost identical predictions, i.e. the estimated VaR and ES basically copy each other. The differences between the lengths of 1 and 30 are negligible as well, however, slight departures can be observed during the financial crisis in the end of 2008. On the other hand, the VaR predictions from models with Gaussian distribution differ substantially from their fitted values. The Gaussian predictions exhibit high sensitivity to the sudden changes in the index prices causing the estimated VaR to drop rapidly. This holds for shocks more than for slight changes but the effect is present for both. Moreover, the predictions based on the longer refit step are even more sensitive.

4.2 Quantile Loss Function and Quantile Dynamic Test

Quantile losses are averaged over the forecasting periods and the preferred model is the one with the lowest average value. The quantile loss (QL) function for time t at risk level α

$$QL_t(\alpha) \equiv (\alpha - d_t)(y_t - VaR_t(\alpha)), \qquad d_t \equiv I\{y_t < VaR_t(\alpha)\},$$

where $I\{\cdot\}$ is an indicator function. Series $d_t, t = 1, \ldots, T$, is called the hitting series and if the model is correctly conditionally covered, d_t should be independently distributed. This is tested by the dynamic quantile (DQ) test which is based on the joint hypothesis that (i) the hitting series are independently distributed, and (ii) the expected proportion of exceedance is equal to the risk level. The null hypothesis of the DQ test can be interpreted as the correct unconditional and conditional coverage and not rejecting the null hypothesis is desired.

The averaged QL functions are calculated for 1% and 5% VaR and the results show that the lowest values belong to the GAS models with the Student's *t* or skew-Student's *t* distribution. Their differences are negligible. On the other hand, the gap between these and the Gaussian model is more profound.

The results show that the QL does not change with the refit step, i.e. if the model is the best-performing one using the refit step of 1 then it also performs best when refit step of 5 or 30 is utilized. However, it does not apply for various confidence levels α, e.g. for FTSE 100 in 2000–2010 period, the model with the Gaussian distribution exhibits a better fit for 5% VaR while the skew-Student's *t* distribution fits the 1% VaR better than the Gaussian one.

Based on the DQ test and the 5% significance level, we cannot reject the null hypothesis of the correct specification for 5% VaR – this applies for all GAS models and indices in each period with the refit step of 1. However, the results for 1% VaR vary. Generally, the DQ test rejects the correct GAS model specification for models with the Gaussian distribution rather than for models with the Student's *t* distribution or skew-Student's *t* distribution. Moreover, for a given refit step in a given time period, the null hypothesis tends to be rejected either for all considered distributions or none of them which is a common and well-documented issue of the DQ test.

4.3 FZ loss function

Despite the QL function for the VaR, there is no such loss function for the ES since it is not an elicitable risk measure. However, the VaR and ES are jointly elicitable using test Fissler and Ziegel loss (FZL) function. Let's assume that VaR and ES are strictly negative and the generated loss differences are homogeneous of degree zero. Then the associated joint loss function FZL for time t at risk level α is formulated as

$$FZL_t^\alpha \equiv \frac{1}{\alpha ES_t^\alpha} d_t(y_t - VaR_t^\alpha) + \frac{VaR^\alpha}{ES_t^\alpha} + \log(-ES_t^\alpha) - 1$$

for the case when $ES_t^\alpha \leq VaR_t^\alpha < 0$. FZL functions are also averaged over the forecasting period and the preferred models are those with lower average values.

FZL values are calculated for all considered indices, periods, refit steps, distributions, and for both 1% and 5% risk levels. The results almost copy the QL function results. Generally, the GAS models with Student's *t* or skew-Student's *t* distribution have the lowest FZL values and the differences between them are negligible. The gap is more profound between the Gaussian and Student's *t* or the Gaussian and the skew-Student's *t*. Furthermore, the gaps between Gaussian and *t* distributions are much higher for 1% VaR than 5% VaR as expected. Overall, the choice of the refit step does not influence the prediction performance of models in terms of the average FZL value. There are only a few exceptions – the differences are negligible and occur usually between the Student's *t* and skew-Student's *t* distribution.

5 Conclusion

In this paper, two risk measures – the Value at Risk and Expected Shortfall – are modeled and predicted by GAS models for four major world stock market indices. The considered GAS models are assessed from different perspectives. First, our results show that GAS models with conditional Student's *t* and skew-Student's *t* distributions perform similarly. Differences in terms of AIC values are negligible and the models are considered to be equally good in all assessed cases for all four indices. The estimated volatility and values of both risk measures differ very slightly. On the contrary, the GAS model utilizing the Gaussian distribution performs worse and leads to extreme values of the estimated volatility and consequently extreme values of the VAR and ES in days of price shocks. Additionally, the t-GAS model is compared to the well-known t-GARCH model. The results are ambiguous in terms of AIC values since examined periods result in no dominance of any model. However, the estimated volatility from t-GARCH reacts more extensively to large values of returns than volatility from the t-GAS model. This is caused by the score of the *t* distribution in GAS dynamics since such large values are considered to be caused by the fat-tailed nature of the data and thus, they are not be fully attributed to increases in the variance. Hence, GAS models might be of a better choice.

The choice of the refit step for rolling windows has a negligible impact on predictions based on GAS models assuming the Student's *t* or skew-Student's *t* distributions. Consequently, the estimated values of VaR and ES are almost indifferent. The choice of the refit step has a higher impact on predictions from the GAS model assuming the Gaussian distribution where the predicted values of VaR and ES tend to be significantly lower than their fitted values in days with price shocks. There is even a noticeable difference among predicted values considering different refit steps.

GAS models with the Student's *t* or skew-Student's *t* distribution outperform the GAS models with the Gaussian distribution in terms of the prediction power. Three different performance measures for predictions are considered: dynamic quantile test, quantile loss function, and Fissler and Ziegel loss function. Our results show that the model ranking is not sensitive to the choice of the refit step, i.e. if the model is the best-performing one when the refit step of 1 is used then it also performs the best when the refit step of 5 or 30 is used. On the other hand, it does not apply for various VaR levels ($\alpha = \{1\%, 5\%\}$), i.e. if the model is the best-performing one when $\alpha = 1\%$ is used, it might not perform the best for $\alpha = 5\%$. However, generally, the differences are negligible between models utilizing the Student's *t* and skew-Student's *t* distributions which both perform better than the Gaussian one.

Acknowledgements

This work was supported by the Internal Grant Agency of University of Economics, Prague under Grant F4/53/2019.

References

[1] Acerbi, C. & Tasche, D. (2002). Expected shortfall: a natural coherent alternative to Value at Risk. *Economic notes*, 31, 379–388.

[2] Ardia, D. & Boudt, K. & Catania, L. (2018). Downside risk evaluation with the R package GAS. *R Journal*, 10, 410–421.

[3] Ardia, D. & Boudt, K.& Catania, L. (2019). Generalized autoregressive score models in R: The GAS package. *Journal of Statistical Software*, 88, 1–28.

[4] Creal, D. & Koopman, S. J. & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5), 777–795.

[5] Delbaen, F. (2002). *Coherent risk measures on general probability spaces*. Springer: Advances in finance and stochastics.

[6] Rockafellar, R. T. & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of banking & finance, Elsevier*, 26, 1443–1471.